Statistical Commission
Fifty-sixth session
4 - 7 March 2025
Item 3(p) of the provisional agenda
Items for discussion and decision: Household surveys

Background document
Available in English only

Handbook of Surveys on Individuals and Households
Foundations and Emerging Approaches
(Draft)

United Nations Statistics Division

Please kindly submit your feedback through the following link:
(https://form.jotform.com/250215155857154)

The form asks for the following information:

- Reviewer information:
  - Name
  - Email
  - Title
  - Organization
  - Country
- Overall comments:
  - Please share your overall comments on the structure and content of the draft Handbook.
  - Please list any areas or topics that you believe the draft Handbook has not covered but should be included.
  - Please list any specific areas where you would like to see more detailed practical guidance.
  - Would you be interested in contributing to the development of case studies? If so, please suggest topics or areas where your expertise could add value.
  - Are there any additional topics or thematic areas that you believe would be beneficial to include in Part II? Please explain why.
- Specific feedback on individual chapters
- Additional advice or feedback not covered in the previous questions that you believe could help improve the draft Handbook
- Whether we can reach out to you for:
  - Reviewing a chapter of the draft Handbook
  - Providing case studies
  - Any other questions we may have

Within this form, you may also submit your comments in a document (MS Word or PDF).

ST/ESA/STAT/SER.F/00

Department of Economic and Social Affairs
Statistics Division

Studies in Methods Series F No. 00

Handbook of Surveys on Individuals and Households

Foundations and Emerging Approaches

Draft submitted to the 56th session of the
United Nations Statistical Commission

United Nations
New York

This page has been intentionally left blank.

TABLE OF CONTENTS

Preface

This is a placeholder for the Preface.

The Preface will be completed in Spring 2025.

Acknowledgements

The *Handbook of Surveys on Individuals and Households: Foundations and Emerging Approaches*, led by the United Nations Statistics Division (UNSD) and guided by the Inter-Secretariat Working Group on Household Surveys (ISWGHS), is a collaborative effort that showcases contributions from a diverse range of experts. It draws on the expertise and experience of professionals from national statistical offices, research institutions, survey firms, and regional and international organizations.

The revision process is led by Haoyi Chen (UNSD), Coordinator of the ISWGHS, under the leadership of Stefan Schweinfest, Director of UNSD, and the co-Chairs of the ISWGHS: Calogero Carletto and Talip Kilic (The World Bank) and Papa Seck and Jessamyn Encarnacion (UN Women). Charles Lau (Gallup), the Lead Technical Editor, has played a pivotal role in technical coordination and project management, supported by Daisy Loayza Murillo (UNSD), who coordinates stakeholder engagement and administrative tasks.

Guidance from a dedicated Steering Committee has played an instrumental role in shaping the Handbook's direction, overseeing technical development, and ensuring effective implementation of the process. The Steering Committee includes representatives from national statistical offices, academia, and survey firms. Members of the Steering Committee include:

- Cilanne Boulet (Statistics Canada)
- Lian Jiajia and Di Jianliang (National Bureau of Statistics of China)
- Leesha Delatie-Budair (Statistical Institute of Jamaica)
- Mauricio Rodríguez Abreu (National Institute of Statistics and Geography, Mexico)
- James Muwonge (Uganda Bureau of Statistics)
- Peter Lynn (University of Essex)
- Frauke Kreuter (Joint Program in Survey Methodology, University of Maryland & LMU Munich)
- Charles Lau (Gallup)

Approximately 80 experts from various institutions contributed to the drafting and review of the handbook, with chapters authored by a broad team of specialists:

- **Introduction**: Haoyi Chen (UNSD), Cilanne Boulet (Statistics Canada), Charles Lau (Gallup), Peter Lynn (University of Essex), Mauricio Rodríguez Abreu (National Institute of Statistics and Geography, Mexico)

- **Chapter 1 (Overview of the Current Survey Landscape)**: Cilanne Boulet (Statistics Canada), Haoyi Chen (UNSD), Charles Lau (Gallup), Peter Lynn (University of Essex)

- **Chapter 2 (Needs, Scope, and Business Case)**: Leesha Delatie-Budair (Statistical Institute of Jamaica), Solly Molayi (Statistics South Africa), Charles Lau (Gallup)

- **Chapter 3 (Survey Management)**: Mauricio Rodríguez Abreu (National Institute of Statistics and Geography, Mexico), Ian O'Sullivan (Office for National Statistics, UK), Octavio Heredia (National Institute of Statistics and Geography, Mexico), Lian Jiajia (National Bureau of Statistics of China), Di Jianliang (National Bureau of Statistics of China), Xiao Yumin (National Bureau of Statistics of China)

- **Chapter 4 (Questionnaire Design, Translation, and Instrumentation)**: Julie de Jong (ICF), Luis Sevilla (NORC at the University of Chicago), Brian Kirchhoff (NORC at the University of Chicago), Brita Dorer (GESIS-Leibniz Institute for the Social Sciences), Peter Takyi Peprah (Ghana Statistical Service)
- **Chapter 5 (Frame and Sampling)**: Juan Pablo Ferreira (National Institute of Statistics, Uruguay), Piotr Jabkowski (Adam Mickiewicz University), Dale A. Rhoda (Biostat Global Consulting)
- **Chapter 6 (Data Collection)**: Elizabeth J. Zechmeister (Vanderbilt University), Elena Cezza (Italian National Institute of Statistics), Gabriella Fazzi (Italian National Institute of Statistics), Muhardi Kahar (Statistics Indonesia), Johannes W. S. Kappelhof (The Netherlands Institute for Social Research/SCP), Michael Robbins (Princeton University)
- **Chapter 7 (Data Processing)**: Fiona O'Riordan (Central Statistics Office, Ireland), Andy Madeley (SwissPeaks Market Research & Data Management)
- **Chapter 8 (Weighting)**: Diego Zardetto (The World Bank), Kevin McGee (The World Bank)
- **Chapter 9 (Analysis)**: Andrés Gutiérrez (Economic Commission for Latin America and the Caribbean), Pedro Luis do Nascimento Silva (Sociedade para o Desenvolvimento da Pesquisa Científica)
- **Chapter 10 (Dissemination)**: Olivier Dupriez (The World Bank), François Fonteneau (OECD)
- **Chapter 11 (Child-Centric Surveys)**: Shane M. Khan (UNICEF), Attila Hancioglu (UNICEF), Turgay Unalan (UNICEF), João Pedro Azevedo (UNICEF)
- **Chapter 12 (Mainstreaming a Gender Perspective)**: Jessamyn Encarnacion (UN Women), Hemalata Ramya Emandi (UN Women), Papa Seck (UN Women), Aurelie Acoca (UN Women), Maria Isabel Cobos Hernandez (UNSD), Iliana Vaca Trigo (UNSD)
- **Chapter 13 (Refugees and Internally Displaced People)**: *to be developed*
- **Chapter 14 (Household Surveys and Education)**: Manos Antoninis (UNESCO), Silvia Montoya (UNESCO)
- **Chapter 15 (Measuring Governance, Public Safety, Justice, and Corruption)**: Alexandre Wilde (UNDP Global Policy Centre for Governance), Arvinn Gadgil (UNDP Global Policy Centre for Governance), Fatma Usheva (UNDP Global Policy Centre for Governance), and Mariana Neves (UNDP Global Policy Centre for Governance), United Nations Office on Drugs and Crime (UNODC)
- **Chapter 16 (Labour Statistics)**: Lara Badre (ILO), Antonio R. Discenza (ILO), Michael Frosch (ILO), Rosina Gammarano (ILO), Vladimir Ganta (ILO), Kieran Walsh (ILO), Samantha Watson (ILO)
- **Chapter 17 (Countries Under Conflict and Post-Conflict)**: *to be developed*
- **Chapter 18 (Small Island Developing States)**: *to be developed*

The Handbook benefited from extensive consultations with national statistical offices, various communities with a focus on special population groups, and expert group meetings, including inaugural discussions in Shenzhen, China (January 2024), and a follow-up meeting in Oslo, Norway (October 2024). Contributions from individual reviewers will be acknowledged in the final draft of the Handbook.

The authors extend their gratitude to Kathleen Beegle, Robert Johnston, Sharon Lohr, Francesca Perucci, and Ibrahim Yansaneh for their instrumental guidance on the revision process. They are also grateful to Cilanne Boulet, not only for her expertise in surveys and her understanding of the challenges faced by a national statistical office but also for her energy, determination, and ability to make things happen. Special thanks go to Lianne Beltran, whose meticulous work on the layout and formatting was essential to bringing this draft to completion.

UNSD extends its deepest gratitude to all contributors for their role in making this endeavour a success. This includes financial support from the Government of China, from members of the Inter-Secretariat Working Group on Household Surveys, and from the Data For Now initiative. We are also grateful to the UNDP Global Policy Centre for Governance for hosting the expert group meeting in Oslo.

GLOSSARY

CHAPTER 1. OVERVIEW OF THE CURRENT SURVEY LANDSCAPE

Adjustment error: Adjustment error occurs when there are differences between the adjusted (weighted) and the true value in the population.

Computer-assisted personal interviewing (CAPI): A mode of data collection where interviewers conduct interviews face-to-face with a respondent, using an electronic data capture system via tablet, laptop, or phone.

Computer-assisted telephone interviewing (CATI): A mode of data collection where interviewers conduct live interviews via telephone, using an electronic data capture system.

Computer-assisted voice interviewing (CAVI): An approach to interviewing that uses technology to create a remote connection between a live respondent and a live interviewer (e.g., via videoconferencing platform). Also referred to as Video Mediated Interviewing (VMI) or Live Video Interviewing (LVI).

Coverage error: Coverage error is a mismatch between the target population and the sampling frame. There are two types of coverage error. Undercoverage occurs when the sampling frame omits units from the target population (e.g., people without phones are excluded from a phone survey). Overcoverage occurs when the sampling frame includes ineligible units from the target population.

Coverage: The extent to which a sampling frame includes all units of the target population, ensuring representativeness and reliability of survey estimates.

Data integration: Data integration refers to the process of combining data from different sources to improve efficiency, cost-effectiveness, granularity, or quality of data. This process can involve combining data from structured and unstructured sources, such as surveys, administrative records, geospatial data, and citizen-generated data.

Generic Statistical Business Process Model (GSBPM): "A framework that describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonised terminology to help statistical organisations to modernise their statistical production processes, as well as to share methods and components." Source: UNECE

Interactive voice response (IVR): An approach to conducting an interview that does not use a live interviewer and, instead, uses technology to guide an interviewee through an automated series of pre-recorded questions while recording their responses either input from keystrokes and/or voice recordings.

Interoperability: To maximize the value of household surveys and facilitate data integration, surveys should be interoperable by design. This means that the survey design and implementation should take into consideration necessary measures to improve the ability of surveys to be integrated with each other and with other data sources.

Measurement error: Measurement error emerges when the respondent's answer deviates from the true value. Measurement error can stem from different sources, including the questionnaire, the respondent, the data capture system, the interview setting, the interviewer, and the mode (Biemer & Lyberg, 2003).

Metadata: Metadata constitute formal and technical documentation of the survey. These may include fieldwork dates, survey mode(s), language(s) of interviews, documentation of weights, the transformation

of data on contact outcomes/dispositions into response rates, information on quality control and related outcomes, and unique identifiers for the survey or participants (e.g., for panel data, so that individuals can be linked across waves).

Mixed mode surveys: A data collection approach in which some respondents may participate in one mode (e.g., CAPI) while others use a different mode (e.g., web surveys).

Multi-mode surveys: A data collection approach in which a respondent uses different modes to collect different subsets of the required data, e.g., administering a self-completion component either within, or subsequent to, a face-to-face interview.

Nonresponse error: Nonresponse error emerges when some sampled units do not participate in the survey (unit nonresponse) or do not answer questions (item nonresponse). Nonresponse error leads to bias if there are systematic differences between respondents and nonrespondents with respect to the outcome of interest.

Non-Sampling error: Errors arising from factors other than the sampling process, such as frame imperfections, nonresponse, and measurement inaccuracies.

Paper self-completion: A mode of data collection where respondents complete a survey on their own (without the assistance of an interviewer) via paper form. Paper self-completion surveys are often mailed to respondents but can also be distributed in-person.

Paradata: Paradata are supplemental data that document the data collection process collected during fieldwork (e.g., length of interview; location data; date, timing, outcome of a contact attempt; deviations from original design, etc.).

Processing error: Errors that occur when data are processed, such as errors in coding, data entry, and tabulation errors, among others.

Sampling error: Sampling error is random error introduced because a survey measures the outcome of interest on a random subset of the target population. It is determined by the sample design, sample size, along with estimation method

Short Message Services (SMS): Messaging that allows the sending and receipt of short text messages across computer / mobile devices.

Target population: The entire set of individuals, households, or entities that the survey intends to represent and to which the results are generalized. Defining the target population clearly is essential to setting the study's scope and eligibility criteria.

Total Survey Error (TSE): A "theoretical framework for optimising surveys by maximising data quality within budgetary constraints" (Biemer, 2010). This Handbook uses TSE as an organizing framework to illustrate the stages of the survey processes and to show how errors can emerge at each stage.

Validity: In the TSE framework, validity refers to the extent to which a survey question accurately operationalizes the construct of interest.

Web surveys: A mode of data collection where respondents complete a survey on their own (without the assistance of an interviewer) via a web browser or app. Also known as "computer-assisted web interviewing" (CAWI).

CHAPTER 2. NEEDS, SCOPE, AND BUSINESS CASE

Business case: The business case is an argument for conducting the survey. It contains information on the justification for the project and the expected value to be obtained from undertaking the survey. A well-prepared business case is crucial for gaining approval and support for surveys, ensuring they are viable, strategically aligned, and well-planned.

External needs: Requirements of stakeholders for what the survey needs to accomplish. External needs represent the desired survey outcomes, to respond to an environmental change (e.g., changes in budget, new methodology) or a new information request (internal or external to NSO). (See also "Internal Needs".)

Internal needs: Requirements that must be met (e.g., availability of funding or human resources) to achieve the desired outcome of the survey. Internal needs represent what is needed to conduct the survey. (See also "External Needs.")

Scope: The Scope consists of both the (1) high-level design of the survey - e.g., target population, mode, conceptual frameworks, topics, etc. and (2) the Statement of Work that lays out specific responsibilities of the implementing organisation.

CHAPTER 4. QUESTIONNAIRE DESIGN, TRANSLATION, AND INSTRUMENTATION

Enabling conditions: A specific set of rules of criteria that determine when a particular question, set of questions, or response options should be activated and available to the respondent.

Interpreting: The conversion of oral speech from one spoken language to another

Skip logic: These are the predefined rules or criteria that determine, based on previous responses, how a respondent advances through the survey.

Source language: The language in which the questionnaire is developed

Survey instrument: The electronic data capture system which the survey questionnaire is programmed into, used to administer the survey

Survey questionnaire: The document with individual survey questions and response options, which are arranged and answered in a predetermined order

Survey specification: Item-by-item instructions to facilitate programming into the data capture system

Target language: The language into which the source language questionnaire is translated

Translatability: The capacity to express the content of source language text into target language text

Translation: The conversion of text from one language into another in written form

TRAPD: A team approach to questionnaire translation, consisting of five steps: (T)ranslation of the source language into the target language(s), (R)eview of the translations, (A)djudication of any inconsistencies in the translation of survey questions to determine the most accurate version, (P)retesting the translations, and (D)ocumentation of the translation process, decisions and the translation team composition

Administrative or Statistical registers: Databases maintained by government agencies or organizations that contain up-to-date information on individuals or households, often used to construct sampling frames.

Area sampling: A sampling method used when person-level or dwelling-level frames are unavailable or when collection is carried out by computer-assisted personal interview (CAPI). The target area is divided into small geographic segments or clusters, and selected primary sampling units (PSUs) are visited to conduct additional rounds of selection and interviews.

Auxiliary information: Additional data related to the primary survey variables, used to improve sampling design efficiency through stratification or calibration.

Census: A comprehensive data collection covering the entire population. It often includes demographic and geographic information useful for creating sampling frames.

Cluster sampling: A sampling technique where the population is divided into clusters, and a sample of clusters is selected. Within each selected cluster, individuals are interviewed.

Complex sample design: A sampling design involving multiple stages, strata, or clusters, requiring adjustments to variance calculations compared to simple random sampling due to its complexity.

Effective sample size: The sample size of a hypothetical survey with simple random sampling that would yield the same margin of error as is observed in an outcome from a complex sample design.

Eligibility unknown: Units for which the eligibility status is unknown, typically due to lack of information or inability to contact them.

Eligible non-respondents (ENR): Units that were selected in the sample and were eligible for inclusion, but for various reasons, they did not respond to the survey. These units may require imputation or weighting adjustments.

Eligible respondents (ER): Units that were selected in the sample, were eligible for and completed the survey, providing the data used for analysis and inferential purposes.

Enumeration area (EA): A geographical segment within a target area, often used in area sampling. EAs are systematically mapped and divided into smaller sections for enumeration.

Finite population correction: Additional factor used to adjust the variance of outcome estimates or sample size calculations when the fraction of the eligible population that is included in the sample exceeds 5%.

Frame units / Sampling units: Elements within the sampling frame (e.g., individuals, households, geographic areas) that represent the potential entities for selection in a sample.

Ineligibility: The status of units within the sample that do not meet the criteria of the target population, leading to overcoverage if those units are included in the frame.

Multiple frames: The use of more than one sampling frame for a survey, often to enhance coverage or precision, managing overlaps to ensure efficient sampling.

Non-response rate: The proportion of sampled units that did not participate in the survey, affecting the representativeness of the sample and requiring adjustments.

Population register: A centralized, government-maintained record of individuals residing within a country, offering high coverage of the target population and often including auxiliary variables.

Probability proportional to size (PPS) sampling: A sampling method where selection probabilities are weighted according to the size of each cluster or unit, helping to ensure the sample reflects the population distribution.

Probability sample: A sample where each individual has a calculable, non-zero chance of being selected.

Response rate: The proportion of sampled units that completed the survey out of all eligible sampled units, impacting the accuracy of survey estimates.

Sampling fraction: The proportion of a group's or subgroup's members that appear in the sample.

Sampling frame: A comprehensive and organized list or database representing the target population from which a sample is drawn, ensuring coverage and representativeness.

Simple random sample (SRS): A sampling method where each individual has an equal chance of being selected and every possible sample of a given size has an equal probability of being chosen.

Stratified sampling: A sampling method where the population is divided into distinct subgroups (strata), and a sample is drawn from each stratum.

Systematic sampling: A sampling method where every nth element is selected from an ordered list, with the first element chosen randomly.

Two-Phase sampling: A sampling method where the population is sampled in two stages: the first phase focuses on obtaining some information, and the second phase collects additional data for selected respondents based on the first-phase responses.

Two-Stage sampling: A method where the sample is selected in two stages: first, a set of clusters is selected, and second, individuals are selected within each chosen cluster.

CHAPTER 6. DATA COLLECTION

Adaptive design: An approach to survey data collection that systematically uses information acquired in prior and, sometimes, in-progress data collection efforts (e.g., on how respondent characteristics predict nonresponse rates) to introduce procedures to increase efficiencies and/or reduce biases in future data collection.

Asynchronous training: An approach to training in which instructors (trainers) and students (trainees) engage at the different times; the instructor prepares material in advance and students participate at different times, often on their own schedule and/or at their own pace

Belmont Report: A document published in 1979 by the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. It outlines three core principles for ethical practices in research that involves human subjects: respect for persons, beneficence, and justice.

Computer audio-recorded interviewing (CARI): An approach to interviewing that uses technology to record parts or all of the interviewer, often for quality control purposes.

Conversational interviewing (CI) approach: A practice in which interviewers are trained and permitted to engage in a more natural conversational style, for example clarifying questions when respondents are having difficulty or may have misunderstood a question; in this approach, uniformity in how questions are asked is sacrificed, to some degree, in order to increase uniformity in the understanding of specialized concepts.

Flags: Thresholds or instances when a survey is considered potentially deviant; used in quality control processes.

General interviewer training (GIT): Training focused on providing foundational information to interviewers about their roles and standard rules for interviewing as well as practical guidance.

Interviewer effects: Behavior that is correlated with interviewers (often, with interviewer traits but also with factors such as workloads [see Wuyts and Loosveldt 2020]), with the potential to increase variance and/or introduce bias into survey outcomes including response rates and substantive responses.

Production monitoring: This involve collecting, assessing, and responding to information regarding the data collection process, which – by comparison to pre-established benchmarks and project materials – allows a statistically informed evaluation and managing of the effort, intervening when necessary to redistribute workloads, document deviations from the original design, and in other ways shape and document the survey process.

Project-specific interviewer training: Training focused on the needs of the specific data collection project – that is, on the study questionnaire, sample, and protocols, and anticipate project-specific challenges

Quality assurance (QA): Establishing processes that create the conditions for success in achieving high quality outcomes.

Quality control (QC): Monitoring, assessing, and responding to information related to QA processes.

Responsive design: An approach that systematically incorporates information acquired during data collection (and potentially before data collection) to support a phased approach to data collection, in which paradata and other information inform decisions in a subsequent stage (e.g., changing mode, incentives, interviewers) to increase efficiency and/or reduce bias in an in-progress data collection effort.

Synchronous training: An approach to training in which instructors (trainers) and students (trainees) engage at the same time, whether in-person or via a distance learning platform.

Train-the-trainer (TTT): A training approach that begins by training a set of trainers, who subsequently train small groups; often used in decentralized training approaches, where trainers are assigned to different groups by geography, language, or another factor.

CHAPTER 8. WEIGHTING

Calibration: Calibration is the process of minimally adjusting the weights so that survey estimates exactly match known population totals. Calibration can improve the precision of survey estimates and help counteract bias.

Design weights: The design weight of a sampled unit is the reciprocal of the unit's sample inclusion probability. They are entirely defined by the sampling design of the survey and, in the absence of non-sampling errors, are sufficient to construct unbiased estimators of population totals.

Inclusion probabilities: Given a sampling design, the inclusion probability of a population unit is the probability that the unit will be included in a random sample selected according to the design. Under probability sampling, every unit in the target population must have a non-zero inclusion probability and it must be possible to calculate the inclusion probability of all selected units.

Response propensity: Response propensity is the probability that a unit in the sample will respond to the survey. Unlike inclusion probabilities, response propensities are unknown and must be estimated from the sample using models. They can be used to counteract nonresponse bias through adjustment of the survey weights.

Survey weights: Survey weights are numerical expansion factors calculated for all surveyed analysis units that allow for valid inference from the sample to the target population. This inference is accomplished through weighted estimators, where the values of the interest variables observed on each analysis unit are multiplied by the survey weight of the unit.

CHAPTER 9. ANALYSIS

Inference: The process of drawing conclusions about a population using data collected on a sample. When using a probability sample, the survey weights allow data users to make estimates that account for the sampling process and produce results that are representative of the population.

CHAPTER 10. DISSEMINATION

Application Programming Interfaces (APIs): Sets of rules and protocols that allow different software applications to communicate with each other. They can be used for survey dissemination by providing a way for survey data to be accessed and integrated into other applications, such as data analysis tools or visualization platforms.

ChatBots: Chatbots are AI-powered virtual assistants that can interact with users through text or voice. They can be used for survey dissemination by providing personalized assistance to respondents, answering questions about the survey.

Descriptive and analytical survey report: This report presents descriptive statistics and an analysis of the survey results, highlighting key findings and their implications for policy or research. It is often the primary reference document for users seeking comprehensive information about the survey outcomes.

Scrollytelling: Scrollytelling is a storytelling technique that combines scrolling motion with visual content and text. It can be used for survey dissemination by creating interactive and engaging presentations of survey data that guide users through the findings in a visually appealing and informative way. Scrollytelling can help to make complex data more accessible and understandable.

Technical survey report: This document provides a detailed account of the survey methodology, including data collection, editing, and analysis processes. It includes lessons learned and a post-mortem evaluation.

This page has been intentionally left blank.

INTRODUCTION

to the

Handbook of Surveys on Individuals and Households

Foundations and Emerging Approaches

Haoyi Chen (United Nations Statistics Division)
Cilanne Boulet (Statistics Canada)
Charles Lau (Gallup)
Peter Lynn (University of Essex)
Mauricio Rodríguez Abreu (National Institute of Statistics and Geography, Mexico)

Introduction

1. This Handbook provides comprehensive guidance on surveys of households and individuals, incorporating both theoretical foundations and emerging approaches. To facilitate the collection high-quality data that can support data-driven policies, this Handbook emphasizes inclusivity, integration within the broader data ecosystem, and the adoption of innovative methodologies to address challenges.

Background: Purpose of the Handbook

2. This Handbook builds on methodological work the United Nations Statistics Division has been conducting since the 1950s. The more recent series started with the *United Nations Handbook on Household Surveys* (United Nations 1984), providing comprehensive guidance on various stages of survey operations. The *National Household Survey Capability Programme* series (1982-1995) covers a range of specialized topics, including questionnaire development, sampling techniques, and surveys focused on income, expenditure, and health ((United Nations 1982a), (United Nations 1982b), (United Nations 1984), (United Nations 1986; 1985), (United Nations 1993)).

3. Later publications, such as *Household Sample Surveys in Developing and Transition Countries* (United Nations 2005) and *Designing Household Survey Samples: Practical Guidelines* (United Nations 2008), offer more advanced, practical insights into survey sampling and implementation, reflecting updated methodologies and addressing challenges unique to emerging statistical systems. Collectively, these publications have been essential in improving the accuracy, efficiency, and comparability of household surveys globally, providing practical tools for generating reliable data on social and economic statistics.

4. Surveys have become increasingly vital in addressing the evolving demands of the modern era, and the last edition of the Handbook—published over 40 years ago in 1984—no longer reflects today's realities. Increasing data demands, advances in survey methodologies, and availability of nontraditional data sources such as administrative data, remote sensing, mobile phone positioning data and citizen data underscore the urgent need for an updated guide for practitioners.

5. This Handbook, a comprehensive update to the 1984 United Nations Handbook on Household Surveys, offers guidance on key principles, necessary steps for each phase of survey design and implementation. The Handbook includes both theoretical foundations and practical solutions, and also includes national experiences and case studies to illustrate practical applications and challenges in diverse contexts.

Audience

6. The Handbook's primary audience consists of National Statistical Offices (NSOs), though it will also be of interest to other stakeholders such as government ministries responsible for data collection through surveys, non-governmental organizations (NGOs), and other organizations with an interest in surveys, private firms conducting surveys, and international and regional organizations operating within countries.

7. The Handbook is designed to support countries with varied statistical infrastructures. It emphasizes key principles, highlights necessary steps for each phase of survey design and implementation and provides national experiences and case studies to for tackling practical challenges instead of being prescriptive.

Features of the Handbook

8. This Handbook carries on the tradition of prior methodological guidance from the United Nations Statistics Division, but also offers four unique features.

9. The first is the coverage of emerging approaches. In addition to theoretical foundations, the Handbook also highlights innovations and new developments. These emerging approaches could refer to methods that are still an active area of research or those that are well-tested but have yet to be broadly adopted in national surveys. In highlighting these approaches, the goal is to motivate national statistical offices and other stakeholders to continue to experiment with emerging approaches and adopt them when they improve quality and enhance time and cost efficiency.

10. The second is inclusion. Traditional survey design assumes that the target population is people living in households. However, this Handbook recognises that it is often desirable for surveys to cover individuals living outside of households if they are also part of the target population that is of interest to the survey organisers. These might include individuals living in institutions, homeless population and other marginalised population groups. Furthermore, traditional survey design assumes that the survey population is sufficiently homogeneous for a single set of survey procedures to suit all. It is increasingly understood that adapted procedures can enable participation by population members who might otherwise be excluded, such as those with various disabilities or who speak different languages. This Handbook points out where procedures may need adaptation in order for a survey to be truly inclusive and, wherever possible, provides references to other resources that may help with the design and implementation of those adaptations.

11. The third is acknowledging surveys as integral components of a broader national data ecosystem. This implies that it is often desirable for surveys to be designed and implemented in coordination with other sources to facilitate data integration and reduce respondent burden. Before starting a survey, a comprehensive assessment on what is already available is necessary. The advantages of building household surveys that are interoperable by design is emphasised throughout the Handbook, as are methods to facilitate this goal.

12. The fourth is that this Handbook will serve as a living document, continuously updated to reflect emerging research and evolving national practices. It will also be made available through an online platform, along with all associated resources.

Structure of the Handbook

13. The Handbook consists of two Parts. Part I of the Handbook provides practical guidance on core tasks in conducting a survey. Part II of the Handbook covers areas that require specific attention for surveys to be inclusive for a specific population groups, subject matter, or specific country contexts.

Part I

14. ***Chapter 1 (Overview of the Current Survey Landscape)*** focuses describes important aspects of the survey landscape that will be recurrent themes throughout the Handbook. Careful reading of the chapter is recommended before referring to other chapters, as this will provide essential context to enable a full appreciation of subsequent guidance. Principles are set out and terms are defined, to provide a common understanding. The first sections of Chapter 1 review fundamental principles of official statistics and quality frameworks for survey data collection. This is followed by discussion of the relative merits of probability and non-probability sampling and of different data collection modes. Issues of inclusivity and respondent burden are discussed next, followed by an introduction to the ideas of paradata and metadata. The final two sections of the chapter address communication and public engagement, as well as integration of data from different sources and how to make surveys interoperable by design.

15. ***Chapter 2 (Needs, Scope, and Business Case)*** outlines the process of planning a survey prior to beginning work. The Chapter covers how to identify needs, considering both external needs (i.e., what outcomes are desired by the survey's stakeholders) and internal needs (i.e., what is required to produce the

survey outcomes). The Chapter also highlights the importance of consultations of key stakeholders and a consideration of alternative approaches (other than a survey). Next, the chapter considers the survey's scope, which includes both major design decisions (e.g., mode, target population, conceptual framework) and the specific Statement of Work. Finally, this Chapter reviews the components and importance of a business case for the survey.

16. Once the survey's needs, scope, and business case are established and resources are allocated, then the process of survey design and implementation (covered in Chapters 3-10) can begin.

17. ***Chapter 3 (Survey Management)*** reviews principles and practices of managing a survey. This chapter explores the essential aspects of survey management, defined as the oversight, direction, and coordination of survey projects. It covers key functions like planning, organizing, leading, and controlling while emphasizing the central role of managing people in this labour-intensive and multidisciplinary process. Survey management bridges planning and technical implementation, ensuring tasks are completed on time, within budget, and to specification. The chapter also highlights how management varies based on organizational context, project type, and other factors, and it defines success through effective stakeholder engagement and balancing time, cost, and quality. Topics such as scheduling, budgeting, risk management, and quality assurance are addressed alongside cross-cutting themes like ethical considerations and stakeholder management, offering a comprehensive guide for successful survey execution.

18. Chapters 4 through 10 each address a core technical task in the survey cycle. These chapters each begin with an overview of the task, goals, and guiding principles. The heart of each chapter is a "How To" guide that provides detailed steps for completing subtasks, identifies risks, and highlights best practices. Chapters conclude with a description of emerging approaches, as well as resources that provide additional guidance.

19. ***Chapter 4 (Questionnaire Design, Translation, and Instrumentation)*** describes the process of questionnaire design. The Chapter begins by reviewing three key conceptual frameworks relevant to questionnaire design: (1) cognitive response process model, (2) questionnaire design, development, evaluation, and testing (QDET); and (3) respondent centred design. Next, the chapter provides guidance and best practices for six subtasks: questionnaire content, technical design for usability, survey specification production, translation, and questionnaire evaluation and testing. This Handbook conceptualizes questionnaire design broadly: it includes not only the process of drafting and translating the content, but also operationalizing the questionnaire in an electronic data capture system (for CAPI and CATI surveys). Given this focus, the chapter also reviews different data capture systems commonly used by NSOs. The chapter concludes with emerging approaches, including machine translation, survey workflow tools, computer-assisted video interviewing (CQVI), web-based cognitive interviewing, and using AI and social media for questionnaire evaluation.

20. ***Chapter 5 (Frame and Sampling)*** starts by describing survey frames. It presented the main types of survey frames typically used for survey of household and individuals, including area frames, dwelling frames, telephone frames and person frames. It gives guidance on strategies for building, maintaining and updating frames, focusing on aspects that are important for producing high-quality frames in practice. The second part of the chapter describes sample design. It introduces the basic types of sampling and shows how they can be used directly or in combination to draw samples that allows for accurate and precise inference to the population of interest in a timely and cost-effective way. The chapter demonstrates the interplay between a survey frame and the sample designs it can support. Practical guidance is provided on sample size calculations and sample allocation methods. More advance sampling methods whose uses are still emerging are also introduced, including methods for reaching hard-to-survey populations.

21. ***Chapter 6 (Data Collection)*** outlines the processes involved in planning and implementing survey data collection, providing the reader with a wealth of suggestions for tried-and-tested methods to ensure that data collection is efficient and effective. The chapter begins by setting out the objectives of data collection. These should be kept in mind while designing each step in the process. Subsequent sections address how to organise a fieldwork team and how to train them – approaches, modes and logistics as well as content – and the organisation of publicity activities that aim to gain support for the survey from the public, specific communities, and/or local authorities. The heart of the chapter is a detailed description of the steps of the data collection process and the decisions that must be made about how best to implement them. This covers strategies for approaching sample members, making contact, gaining co-operation, and carrying out the interview. This is accompanied by sections on production monitoring – standardised observation to track collection processes and facilitate their management – and quality control – to ensure adherence to study protocols and minimise error. The chapter concludes with an introduction to emerging approaches that include the use of data collection modes.

22. ***Chapter 7 (Data Processing)*** focuses on the role of data processing as the bridge between data collection and the production of statistical outputs, ensuring the accuracy, completeness, and usability of survey data for meaningful analysis and reporting. Guided by the General Statistical Business Process Model (GSBPM), it outlines key subtasks such as data extraction, coding, validation, editing, imputation and deriving new variables, while emphasizing principles like relevance, accuracy, coherence, and confidentiality. Best practices, including rigorous documentation, transparent workflows, and the use of automated tools, are highlighted to maintain data integrity and streamline processing. Emerging approaches, such as integrating AI/ML models and transitioning to automated systems, are explored alongside the importance of metadata and quality management. By adhering to these methods and principles, the chapter aims to equip survey practitioners with a robust framework for managing the complexities of data processing in diverse survey contexts.

23. ***Chapter 8 (Weighting)*** describes the process of producing survey weights for probabilistic surveys of households and individuals. The chapter starts with the principles upon which the calculations of weights are based and translates these principles into practical steps. The core of the chapter covers the main weighting steps – design weights, adjustments for non-response, and calibration – and gives details about additional steps such as eligibility adjustments and trimming. Keeping the survey practitioner in mind, the chapter emphasize the development of weights for multistage sample design and includes diagnostics to help monitor quality throughout the weighting process. Finally, the chapter closes by discussing more specialized weighting considerations for panel surveys and pooled samples, along with how the principles and techniques developed for weighting of probability samples can be used in the context of non-probability samples.

24. ***Chapter 9 (Analysis)*** offers comprehensive guidance on analysing household survey data. It begins by introducing the development of an analysis plan, and how that plan can steer both survey design and subsequent secondary data analysis. Based on the foundations of design-based inference, it emphasizes that point estimates should be paired with their associated margins of error and discusses different approaches that can be used for variance estimation. The chapter progresses from basic descriptive statistics—such as totals, means, frequencies, proportions, and quantile—to more intricate analyses, including subgroup comparisons and assessments of variable associations. It also delves into modelling using survey data, discussing the role of weights in estimating linear and logistic regression parameters using data from surveys with complex designs. Additionally, the chapter highlights best practices in data visualization, particularly in representing estimates with their uncertainty measures.

25. ***Chapter 10 (Dissemination)*** offers guiding principles for disseminating survey results, detailing the generation of core products, offering suggestions for effective communication strategies and user engagement, as well as methods for tracking product utilization. It also provides insights into innovative data products and techniques for monitoring use of data products, providing recommendations for developing additional products, and implementing sophisticated usage monitoring to meet evolving user needs.

Part II

26. Part II of the Handbook covers areas that require specific attention for surveys to be inclusive for a specific population group (e.g., gender, children, and refugees and internally-displaced people), or a subject matter (e.g., education, governance, public safety, justice and corruption, and labour), or to be applicable for countries under a specific context (e.g., countries under- or post-conflict and Small Island Developing States).

Population group

27. ***Chapter 11 (Child-Centric Surveys)*** presents best practices for child-centric surveys, defined as surveys that centre the child as the fundamental unit from which all methodological decisions are made. Child-centric surveys generate a package of data on the child and the child's situation. The Chapter begins by discussing how to plan and manage child-centric surveys. Next, the Chapter reviews best practices in questionnaire design, including topics, international standards, questionnaire adaptations for children, and pitfalls in applying adult indicators to children. The Chapter reviews considerations related to sampling, data collection (e.g., consent and assent, ethics, and interviewer criteria), and weighting.

28. ***Chapter 12 (Mainstreaming a Gender Perspective)*** describes best practices in mainstreaming a gender perspective, which is defined as systematically addressing gender issues and mitigating gender-based biases throughout the production and use of official statistics, and across all stages of the data value chain. This overarching principle must be applied to all surveys – not only those focused specifically on women or gender issues but to all types of surveys and data collection efforts. This Chapter reviews how to mainstream a gender perspective throughout the survey cycle, including in planning, survey management, questionnaire design, sampling, data collection, data processing, weighting, analysis, and dissemination.

29. ***Chapter 13 (Refugees and Internally Displaced People)*** will be drafted in 2025.

Subject matter

30. ***Chapter 14 (Household Surveys and Education)*** discusses the use of household surveys as a complementary or alternative source for education statistics, marking its distinctive features. The Chapter highlights key questionnaire design issues, including capturing attendance, attainment, and additional questions. Advanced questionnaire topics include collecting data on technical and vocational training, skills (e.g., literacy and ICT), and education expenditure. The Chapter emphasizes the importance of integrating survey and administrative data for indicators like completion and out-of-school rates.

31. ***Chapter 15 (Measuring Governance, Public Safety, Justice, and Corruption)*** provides guidance and explores best practices for designing and implementing surveys on sensitive governance topics such as corruption, access to justice, political participation, trust, efficacy of service delivery, and victimization. It emphasizes the importance of adopting a broad consultative process with diverse stakeholders to ensure relevance, inclusivity, national ownership and transparency, while adhering to established international standards and methodologies. The Chapter covers key challenges in questionnaire design, sampling, and data collection, offering practical solutions for navigating these complex topics. It also delves into data

dissemination, highlighting the importance of transparent data dissemination protocols to support the publication of politically sensitive data.

32. ***Chapter 16 (Labour Statistics)*** emphasizes the critical role of Labour Force Surveys (LFS) in producing reliable labour statistics aligned with international standards set by the International Labour Organization (ILO). The Chapter highlights key resolutions on labour statistics from the International Conference of Labour Statisticians (ICLS). The Chapter offers best practices for methodology of collecting labour data, including questionnaire design, sampling strategies, weighting adjustments, and the use of administrative data to support LFS. The Chapter also discusses challenges in implementing new standards and highlights the need for clear communication strategies to achieve the information gains intended by the standards and avoid misinterpretation of the breaks.

Countries under a specific context

33. ***Chapter 17 (Countries Under Conflict and Post-Conflict)*** will be drafted in 2025.

34. ***Chapter 18 (Small Island Developing States)*** will be drafted in 2025.

The current state of the Handbook

35. The current draft is prepared, as a background document, for the 56th United Nations Statistical Commission, under agenda item 3(p) on Household Surveys[1]. As of January 2025, the Handbook stands as a largely complete first draft and contains several gaps. These gaps, listed below, will be addressed in 2025 before the Handbook is finalised:

[1] Complete Chapter 3 (Survey Management) and Chapter 7 (Data Processing)

[2] Prepare Chapters 13 (Refugees and Internally Displaced People), 17 (Countries Under Conflict or Post-Conflict), 18 (Small Island Developing States), and other thematic chapters that may be requested by countries.

[3] Expand on the cross-cutting themes of innovation, inclusivity, and interoperability across chapters.

[4] Add more country case studies and national experiences.

[5] Conduct a thorough review and revision to ensure consistency in content, terminology, and key messaging throughout the Handbook. Additionally, efforts will be undertaken to enhance coherence and establish stronger connections between chapters, resulting in a unified and comprehensive final document.

[6] Incorporate feedback on the current draft from the UN Statistical Commission, national statistical offices, and external experts in the respective technical and subject areas through a peer-review process.

The revision process

36. The Handbook revision was led by the United Nations Statistics Division under the guidance of the Inter-Secretariat Working Group on Household Surveys. The process was supported by a governance structure comprising the UN Statistics Division, co-chairs of the Inter-Secretariat Working Group on Household Surveys, the lead technical editor, and a Steering Committee. The Steering Committee includes

[1] https://unstats.un.org/UNSDWebsite/statcom/session_56/documents/2025-19-HouseholdSurveys-E.pdf

representatives from National Statistical Offices across various regions and experts from academia. Approximately 80 experts from national statistical offices, academia, survey consulting firms, and regional and international organizations contributed to the drafting and review process.

37. The revision process benefited greatly from extensive input gathered through multiple rounds of consultations, which addressed challenges related to surveys and the need for guidance in various contexts. These consultations engaged countries across different regions and communities, with a particular focus on thematic areas such as gender, migration, refugees and internally displaced persons (IDPs), and persons with disabilities. Three technical meetings were organised to support the revision process.

Part I

of the

Handbook of Surveys on Individuals and Households

Foundations and Emerging Approaches

United Nations
New York

This page has been intentionally left blank.

CHAPTER 1

OVERVIEW OF THE CURRENT SURVEY LANDSCAPE

Cilanne Boulet (Statistics Canada)
Haoyi Chen (United Nations Statistics Division)
Charles Lau (Gallup)
Peter Lynn (University of Essex)

CHAPTER 1
OVERVIEW OF THE CURRENT SURVEY LANDSCAPE

1. The current survey landscape is marked by both *continuity* and *transformation*. On one hand, the fundamental principles and applications of survey planning, design, and implementation remain stable. For example, sound statistical methodology developed in the 20th century still guides sampling, weighting, and analysis. While quality frameworks have evolved, core principles about survey design and error have stood the test of time. On the other hand, technological changes have sparked transformations in surveys, including an ever-expanding set of data collection modes, ability to link and integrate surveys, and ability to collect and leverage more detailed data about the data collection process. Policymakers are increasingly requesting information about diverse population groups, with higher frequency and for smaller geographic areas. At the same time, national statistical organizations (NSOs) and others are trying to reduce the burden on survey respondents.

2. To set the stage for the subsequent technical chapters in the Handbook, this chapter describes nine aspects of the survey landscape that are critically important to the design and implementation of surveys of household and of individuals. These aspects are cross-cutting issues that will appear across the technical chapters of the Handbook.

- Fundamental Principles of Official Statistics (Section 1.1)
- Quality framework (Section 1.2)
- Probability and non-probability surveys (Section 1.3)
- Modes (Section 1.4)
- Inclusivity (Section 1.5)
- Respondent burden (Section 1.6)
- Paradata and metadata (Section 1.7)
- Communication and public engagement (Section 1.8)
- Data integration and interoperability by design (Section 1.9)

1.1. Fundamental Principles of Official Statistics

3. Surveys of households and individuals are a key tool for producing official statistics about social and demographic topics. Indeed, no alternative data source is available on many topics, and a survey is the only method to gather such information.

4. This Handbook focuses on a range of methodological, operational, and institutional factors when designing and implementing surveys. Subsequent chapters provide established procedures, best practices, and lessons learned to help NSOs produce quality data. Before delving into these topics, it is worthwhile to situate these issues within a larger context: the Fundamental Principles of Official Statistics, FPOS (https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx). These principles provide recommendations for how to govern the production of official statistics, including the need for the legal and institutional framework that guarantees the principals required for statistical work. These principles were first adopted by the United Nations Statistical Commission in 1994 and reaffirmed by the United Nations General Assembly in 2014. Adhering to these principles is a way to bolster trust in the integrity of official statistical systems. Producers of official statistics based on surveys of households and individuals should aim to follow these principles. The United Nations National Quality Assurance Frameworks Manual for Official Statistics, UNNQAFManual (United Nations, 2019) draws attention to how the FPOS and the quality dimensions, presented in the next section, reinforce each other.

5. The ten Fundamental Principles of Official Statistics (FPOS) are:

- **Principle 1.** Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy, and the public with data about the economic, demographic, social, and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.
- **Principle 2.** To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations—including scientific principles and professional ethics—on the methods and procedures for the collection, processing, storage, and presentation of statistical data.
- **Principle 3.** To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods, and procedures of the statistics.
- **Principle 4.** The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.
- **Principle 5.** Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs, and the burden on respondents.
- **Principle 6.** Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.
- **Principle 7.** The laws, regulations, and measures under which the statistical systems operate are to be made public.
- **Principle 8.** Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.
- **Principle 9.** The use by statistical agencies in each country of international concepts, classifications, and methods promotes the consistency and efficiency of statistical systems at all official levels.
- **Principle 10.** Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

6. This Handbook provides practical guidance about how to carry out surveys, which are essential for fulfilling **Principle 1**. In particular, it covers the key considerations and principles needed to collect and prepare (**Principle 2**) and disseminate (**Principle 3** and **6**) official statistics using surveys of households and individuals. The quality framework and cross-cutting themes that are introduced in this chapter and discussed throughout the Handbook align with these Principles (integration and interoperability – **Principle 8**, **9**, and **10**; quality and burden – **Principle 5**).

1.2. Quality framework

7. First and foremost, the objective of this guide is to provide practical guidance for producing quality statistics using surveys that are "fit for purpose"—that is to say, that meet the data users' needs. It is

therefore fundamental to specify what is meant by survey quality, in particular in the context of official statistics. The concept of survey quality is generally recognized to be multifaceted, and its components can be classified along multiple dimensions. While individual countries may express the dimensions of quality slightly differently, there is broad consensus on the concepts they include. In this guide, we will refer to the dimensions as laid out in the UNNQAF Manual (United Nations, 2019). Adopted by the United Nations Statistical Commission in 2019, its recommendations are anchored on the FPOS and encompass the entire national statistical system, including surveys carried out by both national statistical agencies and other producers of official statistics.

8. The United Nations National Quality Assurance Frameworks lay out the dimensions of quality as follows:

- Relevance
- Accuracy and reliability
- Timeliness and punctuality
- Accessibility and clarity
- Coherence and comparability

9. ***Relevance*** is the primary the reason for carrying out a survey. It is the extent to which the statistics satisfy the needs of the users. Relevance is subjective and varies for different user groups. Design decisions are essential for producing relevant statistics. This starts with user consultation which is a key part of defining survey objectives (Chapter 2) and feeds into subsequent design decisions about how to measure concepts (Chapter 4) and which population is surveyed (Chapter 5). Relevance is also a key consideration surrounding dissemination, as addressed in Chapter 10.

10. ***Accuracy*** is the closeness of estimates to the exact or true values that the statistics were intended to measure. It reflects the degree to which the information produced correctly describes the phenomena they as supposed to measure. Accuracy is essential, and producing accurate statistics is important at all survey steps. This is reflected in the fact that the concept is found woven throughout all chapters of this Handbook. Multiple factors can contribute to differences between statistics and the true population values, as described by the Total Survey Error framework described below. Closely related to accuracy, ***reliability*** generally refers to attaining similar results over multiple measurements. In the context of the dimensions of quality, reliability is more specifically the closeness of the initially estimated values to the subsequent estimated values if preliminary figures are disseminated.

11. ***Timeliness*** is the length of time between the end of a reference period (or a date) and the dissemination of the statistics, and ***punctuality*** refers to the time lag between the release date and the target date by which the data or statistics should have been delivered. While statistical agencies should attempt to minimize the delays in making statistics available, balancing the need for timeliness and punctuality with the other dimensions of quality is one of the classic challenges of survey work. As with accuracy, timeliness is a concern throughout the survey process, though it takes on a particular importance in survey management (Chapter 3).

12. ***Accessibility*** refers to the ease and conditions with which statistical information can be obtained, and ***clarity*** is the availability of appropriate documentation relating to the statistics and the additional assistance that producers make available to users. While most relevant to Chapter 9 and 10, which discuss analysis and dissemination, the availability of survey metadata describing the other survey steps contributes directly to clarity. This dimension must be planned for in advance and should not be an afterthought, i.e., left until after data collection and processing are completed.

13. ***Coherence*** is the ability to reliably combine statistics and data sets in different ways and for various uses. Consistency is often used as a synonym for coherence. ***Comparability*** is the extent to which differences in statistics from different geographical areas, non-geographical domains, or time periods can be attributed to differences between the true values of the statistics. It is a key objective when translating questionnaires, and ensuring comparability may require the use of non-literal translations (see Chapter 4). Both coherence and comparability are the result of implementing harmonized processes on a set of surveys or between survey data and data from other sources. Of particular importance to Chapter 4 to 8, this includes the use of common concept definitions and corresponding questions, as well as common collection, data processing, and weighting procedures.

14. Throughout this Handbook, it will be evident that the dimensions of quality are interrelated. These connections include ways in which one dimension reinforces another. For instance, estimates that are not timely are less relevant. On the other hand, there are also trade-offs or compromises between these dimensions. In particular, timeliness and punctuality become more difficult to meet as a result of processes put in place to ensure the other four dimensions. Effective survey management is essential for managing goals that compete with each other at times.

1.2.1. Accuracy: Total Survey Error (TSE) Framework

15. Among the quality dimensions, accuracy is particularly important. Accuracy of a survey statistic refers to the deviation of the estimated value from the true value. The observed value may be different from the true value due to systematic error (bias) and random error (variance).

16. Accuracy is affected by different types of errors that can occur throughout the survey process. Total Survey Error (TSE) is a "theoretical framework for optimising surveys by maximising data quality within budgetary constraints" (Biemer, 2010). This Handbook uses TSE as an organizing framework to illustrate the stages of the survey processes and to show how errors can emerge at each stage.

17. The TSE framework has two main components.

- ***Representation***: The correspondence between survey respondents and the population the survey is intended to represent.
- ***Measurement***: The correspondence between the information recorded from survey respondents and the concept it is intended to measure.

18. **Figure 1.1** shows how representation and measurement are further broken down into stages of the survey process (rectangles) and errors that can emerge between each stage (ovals).

1.2.1.1. Measurement component

19. According to the TSE framework, the measurement component has four stages:

[1] ***Construct***: Identifying and specifying the concepts of interest (i.e., defining the idea or thing to be measured); e.g., unemployment, migration, or a health outcome

[2] ***Measurement***: Designing questions, response options, and data collection systems to operationalize the construct of interest in a way that allows respondents to provide answers

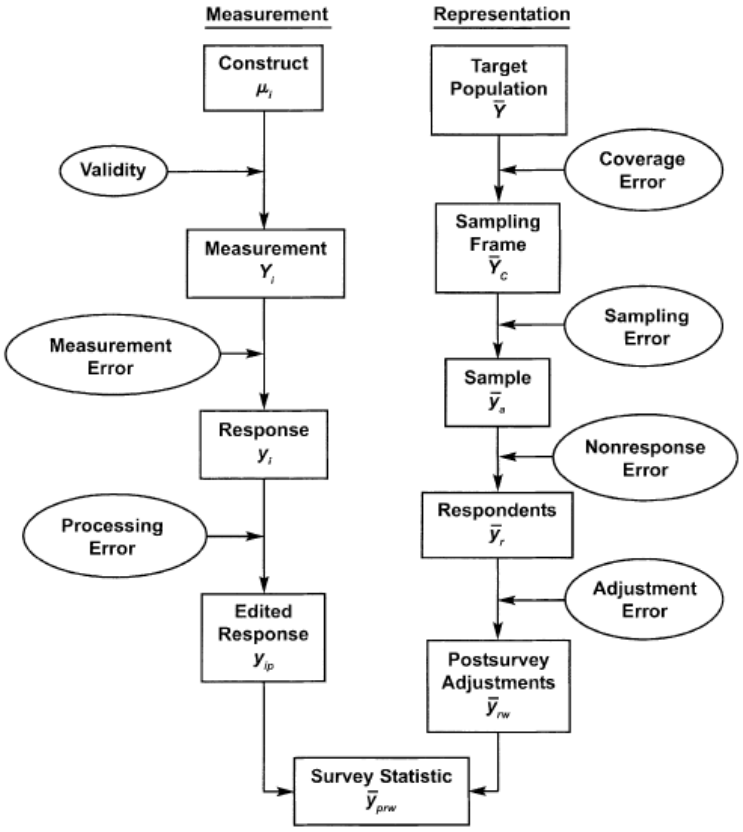
[3] ***Response***: The process of a respondent answering the question. Response can also include the recording of data that are not respondent self-reports, such as anthropometric measurements, interviewer observations, or data from remote sensors, but this Handbook focuses on information self-reported by respondents.

[4] ***Edited response***: Data manipulations that occur after the respondent has answered the question in order to facilitate data analysis

20. Between the above stages, the following errors can emerge:

- ***Problems with validity*** occur when the measurement does not accurately operationalize the construct of interest. Ensuring validity is critical during questionnaire design (see Chapter 4).
- ***Measurement error*** emerges when the respondent's answer deviates from the true value. Measurement error can stem from different sources, including the questionnaire, the respondent, the data capture system, the interview setting, the interviewer, and the mode (Biemer & Lyberg, 2003). Measurement error is discussed in the chapters on questionnaire design (Chapter 4) and data collection (Chapter 6).
- ***Processing error*** refers to errors that occur when data are processed, such as errors in coding, data entry, and tabulation errors, among others (see Chapter 7).

Figure 1.1. Total Survey Error (TSE) Framework (reproduced from Groves, 2009, p. 48). Stages of the survey process are within rectangles; potential errors are shown in ovals.



1.2.1.2. Representation component

21. Within the TSE Framework, there are five stages to the representation process:

[1] ***Target population***: The population that the survey is intended to represent

[2] ***Sampling frame***: The universe of units in the target population, from which a sample can be selected

[3] ***Sample***: The sample of units drawn from the frame

[4] ***Respondents***: Sampled units that are eligible for and complete the survey

[5] ***Postsurvey adjustments***: Adjustments such as weighting and, when applicable, imputation that occur after data collection to make the observed units more closely align with the target population

22. Ideally, all stages of representation are perfectly aligned. In practice, however, errors occur, including the following:

- ***Coverage error*** is a mismatch between the target population and the sampling frame. There are two types of coverage error. Undercoverage occurs when the sampling frame omits units from the target population (e.g., people without phones are excluded from a phone survey). Overcoverage occurs when the sampling frame includes ineligible units from the target population.
- ***Sampling error*** is random error introduced because a survey measures the outcome of interest on a random subset of the target population. It is determined by the sample design, sample size, along with estimation method (see Chapter 5 and 8). Because it is mostly variable in nature, it mainly contributes to and can be quantified via the variance. Some choices of estimators may also introduce systematic error and therefore include a bias component.
- ***Nonresponse error*** emerges when some sampled units do not participate in the survey (unit nonresponse) or do not answer questions (item nonresponse). Nonresponse error leads to bias if there are systematic differences between respondents and nonrespondents with respect to the outcome of interest. If nonresponse error is random, it can increase variance. Survey managers can minimize nonresponse through questionnaire design (Chapter 4) and data collection (Chapter 6). Its effect can also be mitigated via weighting (Chapter 8) and imputation (Chapter 7).
- ***Adjustment error*** refers to differences between the adjusted (weighted) and the true value in the population (see Chapter 8).

23. All errors described above—with the exception of sampling errors—can be defined as "non-sampling errors." Unlike sampling errors, which are known to be a mathematical function of sample design, size, and estimation method, non-sampling errors can be challenging to quantify. During survey design and implementation, it is critical to be aware of, and to mitigate, potential non-sampling errors. Subsequent chapters of the Handbook will refer to sampling and non-sampling errors according to the TSE framework.

1.3. Probability and non-probability surveys

24. For decades, probability sampling and design-based inference, as first laid out by Neyman (1934), have been the cornerstone of official statistics for delivering accurate and reliable population estimates. However, recent years have seen a significant shift, driven by declining response rates, increasing respondent burden, escalating data collection costs, and a growing demand for more timely statistics. Advances in technology and the proliferation of online data sources have further propelled this "wind of

change" (Beaumont, 2020), particularly in developed countries, where many NSOs have pursued modernization initiatives to integrate probability surveys with non-traditional data sources.

25. Amid these developments, non-probability surveys have seen renewed interest. Although not new to survey practice (methods like quota sampling have long produced non-probability samples), non-probability surveys were historically overshadowed by the robust theoretical foundation of probability sampling and design-based inference. Today, non-probability surveys are valued for their affordability, speed, size, and extensive use in online research. One example is the adoption of the Scheveningen Memorandum on Big Data and Official Statistics by the directors general of European NSOs (Eurostat, 2013). Another example is the widespread use of opt-in web panels maintained by commercial survey firms, where individuals agree to participate in online surveys, often motivated by incentives. Despite their practical advantages, deriving accurate and reliable estimates from non-probability surveys remains a persistent challenge. While some NSOs have been cautiously experimenting with non-probability methods, most have not.

26. In probability sampling, every unit has a measurable non-zero probability of inclusion. This allows for the calculation of weights, which in the absence of non-sampling errors, ensure that the estimation process is unbiased. Model-based and model-assisted methods, which depend on stronger assumptions, can subsequently be used as part of the weighting process to minimize the impact of non-sampling errors. These models can correct for nonresponse and coverage bias when it is related to covariates available on the frame or related to variables measured on the survey and available at a population level. Moreover, the use of a probability sample allows quality indicators such as response rates to be calculated. These indicators can be used to monitor and intervene during collection and to inform data users of survey quality.

27. The situation for non-probability sampling is markedly more difficult. The collection process for non-probability samples is often completely unknown and uncontrolled. Weights adjusting for this process must be created solely using model-based methods, even though little may be known about the participation mechanism. This is a fundamentally more challenging task than building weights in a probabilistic setting, especially since there is often less information on which to base the models in a non-probability sample context than there can be in the context of a probability sample. Moreover, the choices required when building model-based methods make them more subjective than design-based methods. As a result, non-probability sampling remains substantially more susceptible to selection bias. To compound these problems, the lack of a frame can make it more difficult to define quality indicators for non-probabilistic surveys.

28. This Handbook aims to provide guidance on how to best implement surveys of individuals and households. *The use of probability sampling remains a key recommendation, and the methods presented are centred around their use within a design-based paradigm.* This paradigm is core to the recommendations of this guide for sampling, weighting, and analysis. Chapter 4 presents how to build frames and sample from them in a probabilistic manner. It reiterates the importance of using probability sampling. Maintaining the probabilistic nature of the sample at later stages of sampling that may happen during collection is also important and is discussed in Chapter 5. Chapter 8 and 9 explain how to prepare weights and use them for design-based inference.

29. Though the use of probability samples within a design-based framework is highly recommended and emphasized in this Handbook, considerations for making better use of non-probability approaches when a probabilistic methodology is not possible are also touched on as emerging approaches. Chapter 8 includes a discussion of the challenges associated with producing weight for non-probabilistic data and an overview of weighting options that have been proposed for this context. This is an active area of research. There is

also growing acknowledgement that non-probabilistic surveys can be improved by thoughtful design decisions, such as including content that can help reduce selection bias (see Chapter 4, questionnaire design). Finally, Chapter 10 makes recommendations surrounding the best practices for informing users of the quality of the surveys. This is particularly important for non-probability surveys given the risks associated with their use.

1.4. Modes

30. Survey data can be collected from respondents in several different ways. This Handbook will devote a lot of attention to face-to-face interviewing, as this remains the predominant form of survey data collection in a large proportion of countries. But telephone interviewing will also be discussed, as will—to a lesser extent—self-completion methods.

31. Face-to-face interviewing can either involve the interviewers writing down the answers by hand (known as pen-and-paper interviewing, ***PAPI***) or the use of laptop computers or tablets to administer the interview (known as computer-assisted personal interviewing, ***CAPI***). Telephone interviewing is usually conducted these days via an electronic questionnaire and is referred to as computer-assisted telephone interviewing (***CATI***). A more recent development is the spread of live video interviewing (***LVI***), where the interviewer communicates remotely with the respondent via video and audio on the respondent's phone or computer. Self-completion surveys can be completed online (***web surveys***), a method which is very popular in some countries, or on paper (***paper self-completion***).

32. Different modes have different strengths and weaknesses. The suitability of a particular data collection mode will depend on the survey context, including local constraints and requirements. PAPI and CAPI are perhaps the most universally applicable modes, as their demands on the local infrastructure or on the capabilities or connectedness of respondents are minimal. However, they are typically also the most resource-intensive modes. The factors that influence the choice of data collection mode are discussed elsewhere in the Handbook, including costs (Chapter 6), availability of contact information on the sampling frame (Chapter 5), population literacy (Chapter 4), availability of computer infrastructure (Chapter 3), sensitivity of the survey topic and risk of social desirability bias (Chapter 4 and 6), risk of mode measurement effects (Chapter 4), and initial post-collection processing steps (Chapter 7).

33. The choice of mode or modes for a survey has fundamental implications for the organization and management of data collection (see Chapter 6), but it also has implications of different kinds for other parts of the survey process, such as questionnaire design (see Chapter 4), sampling (Chapter 5), and data processing (Chapter 7). When choosing the best data collection mode for a survey, several issues should be taken into account (Chapter 6).

34. As different modes have very different strengths, they can be combined in ways that seek to take advantage of the strengths of each mode. Consequently, official surveys in some countries have moved towards implementation of ***mixed-mode*** approaches, in which some respondents may participate in one mode (e.g., CAPI) while others use a different mode (e.g., web surveys). The choice of modes may be offered explicitly to sample members (a concurrent mixed-mode design), or they may first be requested to take part in one mode (typically a less expensive mode such as web), with nonrespondents in that mode subsequently approached in a different mode (typically a more expensive mode that is likely to be successful at improving the response rate, such as PAPI/CAPI). This kind of fieldwork protocol is referred to as a sequential mixed-mode design. There are many different ways of mixing modes in terms of how many and which modes are included, whether they are offered concurrently or sequentially, and what triggers a switch from one mode to the next. Also, the mode(s) in which sample members are first contacted can differ from

the mode in which data is collected. The optimal way of combining modes is likely to depend on a range of contextual factors and constraints, as well as the objectives and content of the survey (de Leeuw, 2005).

35. Some surveys use different modes to collect different subsets of the required data, e.g., administering a self-completion component either within, or subsequent to, a face-to-face interview. We refer to these as ***multi-mode*** surveys rather than mixed-mode, as the modes are not mixed within any survey item: all the responses to a particular question are collected in the same mode.

36. Using a mix of modes can—but will not necessarily—increase participation rates and lead to surveys being more inclusive, compared to single-mode surveys. This benefit can be realized if population members vary in their abilities and preferences to participate in different modes and if those modes can be deployed in a way that encourages participation in the best mode for each sample member. In that situation, different types of people will tend to participate in each mode (a ***mode selection effect***), and the overall sample composition should be a better reflection of the population than those participating in any one mode.

37. However, mixed-mode surveys can also suffer from ***mode measurement effects***, meaning that some respondents would have given a different response if data had been collected in a different mode. This can lead to systematic differences in the data collected in different modes and consequent difficulties in combining them. Sources of mode measurement effects and how to avoid them are discussed in Chapter 4. Some surveys that mix CAPI and self-completion administer the questions most likely to be susceptible to mode measurement effects in a self-completion component within the face-to-face interview in an attempt to avoid mode effects. That said, many survey questions in many contexts—especially simpler, factual ones—seem to be largely immune to mode measurement effects.

38. A final consideration when contemplating the use of a mixed-mode design is that several parts of the survey process will become more complicated. Questionnaire design must simultaneously consider how each question will appear in each mode and seek to minimize differences (see Chapter 4). Sample management systems must allow for the rapid feedback of outcomes in one mode to inform field efforts in another mode; for example, if a respondent completes a web questionnaire while interviewers are in the field attempting face-to-face interviews, the relevant interviewer(s) should be informed, so that they do not visit the address to request an interview. Data management systems must be designed to accept and process data from each mode within one system that produces a single standardized data set, despite likely differences between modes in some aspects of the initial data structure and format (see Chapter 7).

39. The optimal mode, or combination of modes, for a survey will therefore depend on a range of factors.

1.5. Inclusivity

40. Inclusivity within a survey context has three main components: the data to be collected (inclusive data), the methods used to collect the data (inclusive data collection), and the analysis and statistics produced using the data (inclusive statistics). All are important, and collectively contribute to the objective of ensuring that evidence used for social policy making represents and reflects the relevant circumstances of everyone (e.g., the UK goal that "everyone [in the UK] counts and is counted, and no one is left behind"; UK Statistics Authority, 2020) and can therefore support transformative policies (e.g., in line with the Agenda for Sustainable Development promise to Leave No One Behind; UNSDG, 2019).

41. ***Inclusive data*** are data that capture important dimensions of people's lives that should be taken into account in social policy. An example is the need to collect data on characteristics such as gender, religion, and ethnic origin using classifications that reflect the range of categories with which individuals might identify.

42. ***Inclusive data collection*** means using survey methods that are designed to be inclusive of the needs of all members of the population. No one should be excluded from survey participation because the survey method does not enable them to take part.

43. ***Inclusive statistics*** are statistics that recognize and reflect the existence and relevance of all population subgroups whom social policy must serve. For example, official statistics publications should include analysis by gender, religion, and ethnic origin.

44. The notion of inclusivity most pertinent to this Handbook is that of inclusive data collection. This in itself is desirable for moral and ethical reasons (each member of society should have an equal right to have their voice heard and to have their needs considered) as well as for statistical reasons, as it is a means to achieve more representative survey samples (see Chapter 5). But it should be noted that it is also an essential ingredient in a system that aims to produce inclusive data and inclusive statistics.

45. Representative samples are important in order to ensure the accuracy of survey-based population statistics, while the production of statistics on relevant subgroups (inclusive statistics) requires adequate survey participation by members of those subgroups. As a result, it is imperative to ensure that survey sampling and collection activities address issues that can hinder the inclusion of subgroups that are at risk of exclusion. Beyond these two steps, seeking and including the input from a wide variety of groups while defining objectives, building and testing questionnaires, and preparing analysis and dissemination is a way to make surveys more relevant and build trust in statistical agencies and their outputs.

46. To survey data collection inclusively, it is first necessary to take stock of the subgroups likely to be at risk of exclusion. These will vary between countries and contexts, but they will also depend on survey design. For example, those without internet access or functionally unable to interact with a computer are at risk of exclusion from a web survey, but not necessarily from a CAPI survey. There are four main reasons why subgroups may be at risk for exclusion: exclusion by design, sampling frame undercoverage, barriers to participation, and unwillingness to participate. Each of these four should be considered in turn in order to identify the at-risk subgroups and to consider how the risk of exclusion could be reduced.

1.5.1. Exclusion by design

47. Household surveys typically define the target population as those living in private households, and yet the resulting statistics are often interpreted as reflecting the total population and are used to inform policies that affect the total population. Depending on the aims of survey and the extent and nature of the non-household population, the appropriateness of the population definition may therefore sometimes be questioned. In many countries, the non-household population includes people residing in institutions such as prisons, military barracks, and nursing homes, as well as those who are homeless or do not have a fixed address. Belonging to these non-household populations is often a transitory state, not a permanent characteristic of an individual, so the consequence is that people of the kind who spend time in those institutions will be under-represented in household surveys.

1.5.2. Sampling frame undercoverage

48. The frame or method used to select the survey sample may be imperfect, leading to some population members having no chance of inclusion. This causes disproportionate exclusion of subgroups if those missing from the frame are systematically different from others (see Chapter 5).

1.5.3. Barriers to participation

49. There are a variety of reasons why some people may be unable to take part in a survey, even if in principle they would be willing to do so. These include access barriers, language differences, and a range of disabilities and impairments, including deafness, blindness, speech disabilities, cognitive disabilities, motor impairment, neurodivergence, and mental health issues. Access barriers for a face-to-face survey typically consist of locked communal entrances to a set of dwellings, such as gated communities or certain types of apartment blocks.

50. The relative importance of disabilities and limitations depends on the survey mode. A blind person may be able to take part in a face-to-face interview but not in a self-completion survey, while the opposite may be true of a deaf person, for example. Surveys can strive to be inclusive by reducing or removing the barriers to participation. For example, language barriers can be removed by offering questionnaires or interviews in multiple languages (see Chapter 4 for discussion of issues in translating questionnaires); materials in braille and adapted interviews can be offered to blind people (interviewer-administered without reliance on visual presentation of materials); web questionnaires can be designed for compatibility with text-to-speech apps, and so on. Many of the potential adaptations that can be introduced are costly, and a balance may need to be struck between the desire to be completely inclusive and the likely small impact on accuracy of statistics if very small numbers of sample members are affected. Barriers can be lowered simply by putting the respondent at the heart of survey design (Wilson, 2021), and offering alternative modes of participation can be a good way of making a survey more inclusive.

51. ***Unwillingness to participate***: Some people are less willing than others to participate in surveys. If those who do not participate differ systematically from those who do, regardless of the extent of nonresponse, nonresponse bias will result, meaning that subgroups who share characteristics with the nonrespondents will be under-represented. For this reason, surveys increasingly use targeted (Lynn, 2017) or adaptive (Schouten, 2018) designs that put extra efforts into attempting to engage subgroups that would otherwise be less willing to participate. Targeted design features could include introductory letters or messaging that varies between subgroups (Lynn, 2016), different kinds of incentives, or extra field efforts. Other ways of improving the participation of key subgroups include involving them in the design process to ensure that questions and terminology are relevant and understood and the matching of interviewers to respondents in terms of relevant characteristics such as ethnicity, gender, language, or religion. Methods to maximize response and to achieve balance in response propensities between subgroups are discussed further in Chapter 6.

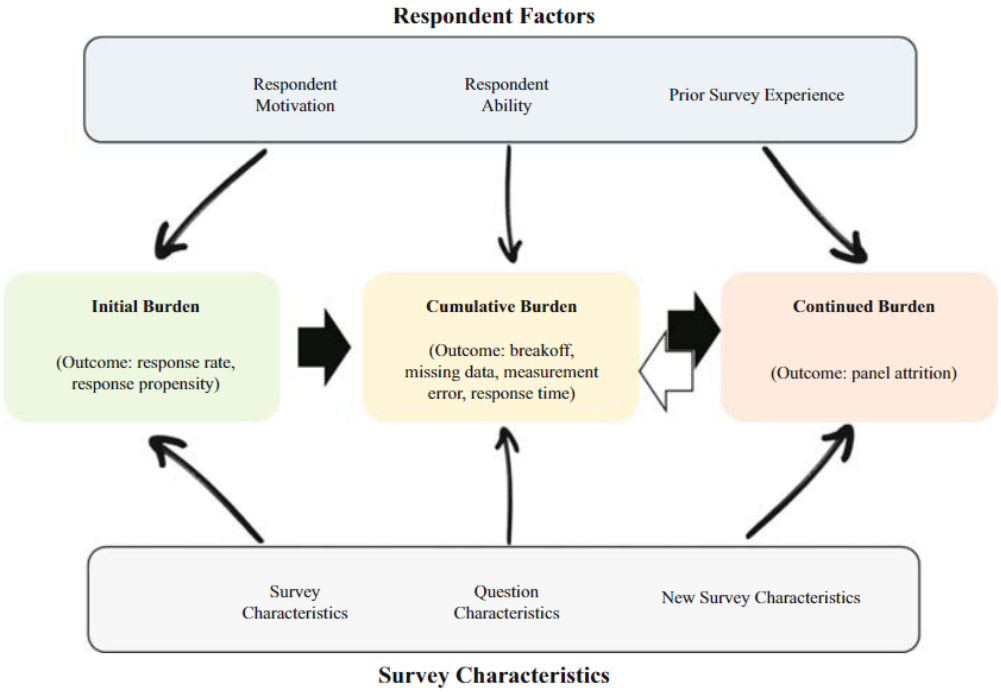
52. In addition to considering all the reasons why subgroups may be at risk of exclusion from a survey, good practice to minimize the risk of exclusion also includes that the survey management and broader implementation team should itself be inclusive. This will help to ensure that the views and experiences of relevant groups inform the survey design. Similarly, an inclusive interviewing team is an asset during collection.

1.6. Respondent burden

53. Respondent burden refers to the actual or subjective perception of the amount of effort respondents devote to completing a survey. Survey researchers have long been concerned with respondent burden (e.g., Sharp and Frankel, 1983). However, attention to respondent burden has grown substantially in recent years due to low response rates and rising costs, amid increasing demand for data. On the micro level, burden may cause a respondent to break off a survey or provide poor-quality responses, leading to nonresponse or measurement error. On the macro level, surveying certain population groups (known as “over-surveying”) can lead to citizens being disengaged or distrustful of government institutions.

54. Yan and Williams offer a conceptual framework for respondent burden (Yan and Williams, 2022), highlighting the impact of respondent factors (e.g., motivation, ability, prior survey experience) and survey characteristics (e.g., mode, topic, question characteristics) on respondent burden (**Figure 1.2**). This framework highlights errors that can emerge at different phases of the survey process: respondents who perceive an undue amount of burden may choose not to respond to an initial survey request (unit nonresponse), break off the survey, choose to say "don't know" or refuse to answer questions (item nonresponse), or provide poor-quality responses (measurement error). Panel respondents may also choose not to participate in subsequent waves of the study.

Figure 1.2. Conceptual framework of respondent burden. Reproduced (Yan, 2022).



55. Some countries have legal frameworks that seek to reduce respondent burden. For example, government agencies in the United States of America must apply for clearance from the Office of Management and Budget (OMB) and justify the need for the collection of surveys (US Office of Personnel Management, 2011). In Europe, Principle 9 of the European Statistics Code of Practice concerns "Non-Excessive Burden on Respondents," requiring that primary data should only be collected if it is "absolutely necessary", if the burden is shared across populations, and if data linkage and integration is used to minimize respondent burden (European Union, 2017).

56. Minimising respondent burden while meeting stakeholder requests for data is an increasing challenge in household surveys. While questionnaire design may be the most apparent method to reduce respondent burden (e.g., questionnaire length, structure, complexity), respondent burden can also be minimized through efficient sampling (see Chapter 5), data collection protocols (see Chapter 6), and data integration and interoperability, among other methods.

1.7. Paradata and metadata

57. In addition to respondent-provided data (from the interview) , household surveys also produce two other types of data: paradata and metadata. These forms of data are not provided by respondents, but are produced by electronic data capture systems and the team conducting the survey. Paradata and metadata are critical for documenting, monitoring, and improving household surveys.

1.7.1. Paradata

58. Paradata are supplemental data that document the data collection process collected during fieldwork (Kreuter, Couper, & Lyberg 2010). Technology has greatly expanded the amount of paradata that, via advance planning and programming, can be passively collected during computer-assisted interviewing. Some paradata can also be recorded by members of the data collection team (e.g., interviewer observations about the household or community). Paradata apply to all modes, but CAPI and CATI surveys produce the most paradata. (For paper self-completion surveys, the questionnaire needs to include instructions for recording paradata such as interview date and start/stop times.)

59. Examples of paradata include contact attempts (e.g., number of times a sample unit is contacted, days and times of contacts, outcomes or dispositions of contact attempts), length (e.g., overall length of interview, length of specific questions or sections), reasons respondents provide for refusals, interviewer observations about a respondent or the housing unit, location data (e.g., GPS coordinates of household), interviewer's keystroke data, and respondent mouse clicks in web surveys, among others (Cohen and Warner 2021; West, 2011). Paradata also include deviations from the original design (e.g., a sample substitution).

60. Paradata can be used for numerous applications (Kreuter, 2013). For example, paradata can be used during fieldwork to identify, monitor, and resolve unit nonresponse that can lead to bias (Kreuter and Olson, 2013). Paradata can also shed light on respondent–interviewer interactions, usability issues, and measurement errors (Yan and Olson, 2013). Paradata are critical for quality control, e.g., to identify interviewers with disproportionately high refusal rates (see Section 6.8 Quality Control in Chapter 6). Paradata can also be used to gain efficiencies through adaptative survey design (Lepkowski, 2010). These are but a few examples of the many uses of paradata: researchers are actively exploring new and innovative techniques to analyse paradata to improve survey quality and efficiency.

1.7.2. Metadata

61. Metadata constitute formal and technical documentation of the survey (Kreuter 2013; Schenk and Reuß 2024). These may include fieldwork dates, survey mode(s), language(s) of interviews, documentation of weights, the transformation of data on contact outcomes/dispositions into response rates, information on quality control and related outcomes, and unique identifiers for the survey or participants (e.g., for panel data, so that individuals can be linked across waves). The UNNQAF Manual for Official Statistics (United Nations, 2019) also provides recommendations for assessing, managing, and reporting survey quality using tools such as quality indicators, reports, assessments, and audits. Using these tools calls for the recording of metadata throughout the survey process.

62. Metadata are sometimes based on paradata: an example is nonresponse rates, which are a type of metadata that are calculated by collecting auxiliary data on successful and unsuccessful contact attempts. Paradata and metadata can be used to calculate key performance indicators (KPIs) that can be monitored during fieldwork for quality control.

63. They can also be used to define important quality indicators that should accompany the dissemination of survey results to inform users about survey quality. Indeed, transparency in research requires comprehensive and accurate metadata. For an example of the types of metadata required to meet the standards in the profession, see AAPOR's Transparency Initiative (https://aapor.org/standards-and-ethics/transparency-initiative/) and [[after UNSC, we will add a global source]].

1.8. Communication and public engagement

64. [[This section will be added in spring 2025.]]

1.9. Data integration and interoperability by design

1.9.1. Integrating data from different sources

65. Data integration refers to the process of combining data from different sources to improve efficiency, cost-effectiveness, granularity, or quality of data. This process can involve combining data from structured and unstructured sources, such as surveys, administrative records, geospatial data, and citizen-generated data.

66. A brief summary of different types of data integration are provided below. These examples highlight that data integration can be useful at many steps of the survey process.

1.9.1.1. Integration of different frames

67. Integrating different frames from which to draw a sample is generally done to serve two primary objectives. The first, the multiple frame approach, is used to enhance frame coverage (see Chapter 5 on sampling and frames for more detail). For instance, combining a mobile phone frame with a landline telephone frame provides more complete household coverage. Similarly, an outdated housing unit frame can be supplemented with a frame of new constructions to ensure newly built housing units are included.

68. The second objective is to improve sampling efficiency by adding supplementary auxiliary information to an existing frame that can be used to stratify samples. For example, the Frames Project[1] in the United States integrates the master address frame with the business register, job lists, and the demographic frame. In this case, demographic, social, and economic variables from the demographic frame can be used to create strata, enhancing the sampling process.

1.9.1.2. Record linkage

69. Record linkage is the process of finding and linking records that refer to the same entity across different data sources. There are mainly two types of record linkage, deterministic and probabilistic. Deterministic linkage uses unique identifiers across different data sources to link records, while probabilistic linkage works when there are no unique identifiers or the information is incomplete or subject to error. In this latter case, each pair of records from the two data sources is assigned a probability of matching.

[1] Frames Program update, 2024. https://www2.census.gov/about/partners/cac/sac/meetings/2024-03/presentation-frames-progam-update.pdf

70. Record linkage can serve multiple purposes: reducing survey burdens and improving data quality; understanding the accuracy of the data sources; and bridging data gaps and supporting more in-depth analysis. Using administrative records such as tax records for data on income in the Canada Income Survey[2] has been shown to help reduce response burden and improve measurement quality. Similarly these same benefits have also been observed when job and social security records have been used to derive information on economic activities in the European Union coordinated surveys that collect data on income and living conditions (EU-SILC (Eurostat, 2013)). The accuracy of Medicaid reporting was assessed through linking the American Community Survey and administrative records.[3] Data linkage was also used to assess connections between housing assistance, health insurance coverage, and unmet medical needs.[4]

1.9.1.3. Calibration

71. Calibration is a statistical technique used to adjust the weights of sampled units, so that survey estimates align with known population totals. Here, the known population totals are from a different, more reliable data source such as census or administrative data. Calibration is a very common practice in household surveys (more details available in Chapter 8) and can be used to correct for nonresponse and coverage errors, as well as ensure coherence between data sources.[5]

1.9.1.4. Small area estimation (SAE)

72. Small area estimation (SAE) is a method that combines two data sources together through a modelling procedure. One data source, typically a survey, collects the outcome variable of interest for the data integration project (e.g., poverty), as well as a wealth of auxiliary information (such as age, sex, education, and employment), but it has less precision than desired due to sample-size constraints. The other data source typically has more data available at more granular levels, but it does not collect data for the outcome variable. The source could be a population census, administrative records, remote sensing/geospatial information, or another survey that has a larger sample size. The SAE method[6] combines these two data sources with an assumption that individuals or households sharing the same characteristics from the two data sources lead to the same outcome. With SAE, accuracy and granularity (both spatial and temporal) of estimates can be improved. One notable example is the long-standing Small Area Income and Poverty Estimates (SAIPE) Program led by the US Census Bureau. SAIPE integrates data from two surveys in the US (American Community Survey and the Current Population Survey) with a number of administrative records such as the tax record and the participant data of the Supplemental Nutrition

---

[2] Statistics Canada. Canadian income survey products [Internet]. 2014. Available from: https://www150.statcan.gc.ca/n1/pub/75-513-x/75-513-x2014001-eng.htm

3 Boudreaux M, Noon JM, Fried B, Pascale J. Medicaid expansion and the Medicaid undercount in the American Community Survey. Health Serv Res. 2019 Dec;54(6):1263-1272. doi: 10.1111/1475-6773.13213. Epub 2019 Oct 10. PMID: 31602631; PMCID: PMC6863241.

4 Simon, A. E., Fenelon, A., Helms, V., Lloyd, P. C., & Rossen, L. M. (2017). HUD housing assistance associated with lower uninsurance rates and unmet medical need. Health Affairs, 36(6), 1016–1023.
National Academies of Sciences, Engineering, and Medicine. 2023. Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources. Washington, DC: The National Academies Press. https://doi.org/10.17226/26804.

[5] Brakel, Jan & Bethlehem, Jelke. (2008). Model-based estimation for official statistics.

[6] For an overview of literatures and methods for small area estimation, consult with the UNSD Toolkit on Using Small Area Estimation for SDGs, available https://unstats.un.org/wiki/display/SAE4SDG/SAE4SDG

Assistance Program (SNAP) to produce annual income and poverty statistics for all school districts, counties, and states.[7]

1.9.2. Maximizing interoperability

73. To maximize the value of household surveys and facilitate data integration, surveys should be interoperable by design. This means that the survey design and implementation should take into consideration necessary measures to improve the ability of surveys to be integrated with each other and with other data sources. Here is a list of ways in which surveys can be made to be more interoperable.[8]

1.9.2.1. Harmonization

74. Harmonization is an important initial step to ensure successful data integration. For surveys to be interoperable by design, the concepts, definitions, and classifications in surveys need to be consistent with other data sources. This can be managed through a metadata system (e.g., Statistics Canada). Another important element to consider is to have the surveys adopt a common set of geographies that ensure that statistical data is geospatially enabled in a consistent manner, whether in gridded form or using administrative or statistical boundaries.[9] Common geographies enable basic statistical reporting, geostatistical analyses, and visualization at different scales (such as at local, sub-national, national, regional, and global, as well as institutional), where the resulting outputs can be compared and assessed on a consistent basis. These geographies also provide a mechanism to enable the management of privacy and confidential statistical and geospatial data outputs.

1.9.2.2. Georeferenced household surveys

75. This refers to linking the data collected in household surveys to specific geographic locations by assigning geographic coordinates (e.g., latitude and longitude). Georeferencing can be very helpful for validating models that combine survey data with satellite imagery and for using geospatial data to generate precise estimates of poverty, asset wealth, and agricultural outcomes at a high spatial resolution. For example, one study integrated publicly available satellite imagery with georeferenced data from Demographic and Health Surveys at the enumeration area level to predict asset wealth in 20,000 African villages.[10] Georeferenced surveys also allow the integration of survey data with environmental information from administrative sources.

1.9.2.3. Common auxiliary variables

76. To enable integration with other data sources, common "auxiliary" variables may also be collected in surveys. As described above, the SAE method relies on a set of common "auxiliary" variables for the model to work effectively. For example, age, sex, education, and type of residence (urban or rural) might be important predictors for poverty or employment. If the survey being conducted is going to be integrated

[7] https://www.census.gov/programs-surveys/saipe/about.html

[8] ISWGHS blog – Chen & Perucci (2023). Data integration – what is 'interoperability-by-design' for household surveys? Available https://unstats.un.org/iswghs/BlogDetails/what-is-interoperability-by-design-for-household-surveys

[9] United Nations Statistics Division, 2019. The Global Statistical Geospatial Framework, available https://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf

[10] Yeh C, Perez A, Driscoll A, Azzari G, Tang Z, Lobell D, et al. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nat Commun [Internet]. (2020) Dec; 11: (1): 2583. Available from: http://www.nature.com/articles/s41467-020-16185-w

with other data sources (censuses or administrative records) for poverty statistics, then these variables should be collected in the survey; and ideally, in comparable definitions and classifications. A study by the Pew Research Center[11] demonstrated that selecting the right variables for weighting to correct biases in an online non-probabilistic survey is more important than selecting the correct statistical method. Thus, for successful data integration, it is important to identify and collect a common set of effective predictive variables in data sources.

1.9.2.4. Unique identifiers

77. Obtaining unique identifiers in surveys for individuals, firms, and facilities is a recommendation that requires careful consideration. On the one hand, unique identifiers facilitate unit-record linkages of household surveys with administrative data sources. In the Frames project of the United States, one of the key recommendations was to have unique identifiers in all data sources used for constructing frames to support integration. On the other hand, concerns regarding privacy may discourage survey participation when respondents are asked to provide unique administrative identifiers.

1.9.3. Legal, policy, and technical aspects of data integration

78. To facilitate data integration, three broad areas are important: legal, policy, and technical. For the legal aspect, a national statistical law or other laws/policies need to be in place to support data integration and to take into consideration ethical and human-rights concerns. These include respecting an individual's privacy and confidentiality; obtaining appropriate consent for using and integrating data; providing individuals with the right to understand how their data are being used; and ensuring that the information is not used to discriminate against or target certain population groups. These legal considerations align with **Principle 2** and **6** of the FPOS (see Section 1.0), which emphasizes scientific principles and professional ethics, as well as the protection of individual data and the confidentiality of information provided for statistical purposes.

79. When considering policy, the following are important: data governance and coordination policies to support data access, while protecting confidentiality; adopting open data and data sharing policies within the national statistical system to facilitate data access and integration; adopting standards and harmonization policies including statistical standards (concepts and definitions), common data formats, and metadata policies; and adopting ethical and privacy policies. These align with **Principle 5** of the FPOS on data for statistical purposes to be drawn from all types of sources and with **Principle 8** on the importance of coordination among all statistical agencies within countries.

80. Technical infrastructure and practices are essential for efficient and secure data integration. These include developing and maintaining interoperable data infrastructure to facilitate the seamless integration of data; using encryption, secure storage, and access control to protect sensitive data during and after integration; adopting robust methods for data linkage, cleaning, and quality assurance; and maintaining a well-documented and real-time metadata system. These technical infrastructure considerations correspond to a number of principles of the FPOS, including the use of professional standards, methods and procedures (**Principle 2**), confidentiality (**Principle 6**), and coordination (**Principle 8**).

[11] https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/

1.10. References

Beaumont, Jean-François. 2020. “Are probability surveys bound to disappear for the production of official statistics?” *Survey Methodology* 46(1), 1-29. https://www150.statcan.gc.ca/n1/pub/12-001-x/2020001/article/00001-eng.htm

Biemer, Paul P. 2010. “Total Survey Error: Design, Implementation, and Evaluation.” *Public Opinion Quarterly* 74(5): 817–848, https://doi.org/10.1093/poq/nfq058

Biemer, Paul P., Lars E. Lyberg. 2003. *Introduction to Survey Quality.* Hoboken, NJ: John Wiley & Sons.

Carletto, Calogero, Haoyi Chen, Talip Kilic, and Francesca Perucci. 2022. “Positioning household surveys for the next decade.” *Statistical Journal of the IAOS* 38(1): 1-24. https://doi.org/10.3233/SJI-220042

Cohen, Mollie J., Zach Warner. 2021. "How to Get Better Survey Data More Efficiently." *Political Analysis* 29(2): 121-138. http://dx.doi.org/10.1017/pan.2020.20

De Leeuw, Edith D. 2005. “To Mix or Not to Mix Data Collection Modes in Surveys.” *Journal of Official Statistics* 21(2): 233–255.

European Statistical System Committee. 2018. European Statistics Code of Practice for the National Statistical Authorities and Eurostat (EU Statistical Authority), 16th November 2017. Retrieved from https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7?t=1528447068000

Eurostat. 2013. The use of registers in the context of EU-SILC: Challenges and opportunities. Eds: Markus Jäntti, Veli-Matti Törmälehto, and Eric Marlier. Retrieved from https://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF

Groves, Robert M.., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau. 2009. *Survey Methodology.* Hoboken, NJ: John Wiley & Sons.

Kreuter, Frauke. 2013. *Improving Surveys with Paradata: Analytic Uses of Process Information.* Hoboken, NJ: John Wiley & Sons.

Kreuter, Frauke., Mick P. Couper, Lars E. Lyberg. 2010. “The Use of Paradata to Monitor and Manage Survey Data Collection.” *Proceedings of the Joint Statistical Meetings, American Statistical Association*, 282–96. Alexandria: ASA.

Kreuter, Frauke, Kristen Olson. 2013. “Paradata for Nonresponse Error Investigation.” Chapter 2 in *Improving Surveys with Paradata: Analytic Uses of Process Information.* 13-42. Hoboken, NJ: John Wiley & Sons.

Lepkowski, James, William Axinn, Nicole Kirgis, Brady T. West, Shonda Kruger Ndiaye, William Mosher, and Robert M. Groves. 2010. “Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection.” *NSFG Survey Methodology Working Papers*, No. 10-012. https://www.researchgate.net/profile/James-Lepkowski/publication/236625287_Use_of_Paradata_in_a_Responsive_Design_Framework_to_Manage_a_Field_Data_Collection/links/543509d40cf2bf1f1f27e8bb/Use-of-Paradata-in-a-Responsive-Design-Framework-to-Manage-a-Field-Data-Collection.pdf

Lohr, Sharon L., and Trivellore E. Raghunathan. 2017. “Combining Survey Data with Other Data Sources.” *Statistical Science*, 32(2): 293-312. http://dx.doi.org/10.1214/16-STS584

Lynn, Peter. 2016. ‘Targeted appeals for participation in letters to panel survey members’. *Public Opinion Quarterly,* 80(3): 771-782. https://doi.org/10.1093/poq/nfw024.

Lynn, Peter. 2017. “From standardized to targeted survey procedures for tackling non-response and attrition.” *Survey Research Methods*, 11(1): 93-103. http://dx.doi.org/10.18148/srm/2017.v11i1.6734

Neyman, Jerzy. 1934. “On the Two Diferent Aspects of the Representative Method: The Method of Stratifed Sampling and the Method of Purposive Selection.” *Journal of the Royal Statistical Society* 97: 558–625.

Schenk, Patrick Oliver, and Simone Reuß. 2024. "Paradata in surveys." In Perspectives on Paradata: Research and Practice of Documenting Process Knowledge, pp. 15-43. Cham: Springer International Publishing, 2024.

Schouten, Barry, Andy Peytchev, and James Wagner. 2017. *Adaptive Survey Design*. New York: Chapman and Hall.

Sharp, Laure M., Joanne Frankel, “Respondent Burden: A Test of Some Common Assumptions,” *Public Opinion Quarterly*, Volume 47, Issue 1, SPRING 1983, Pages 36-53. https://doi.org/10.1086/268765

United Kingdom (UK) Statistics Authority. 2020. Statistics for the Public Good: UK Statistics Authority Five-year strategy 2020 to 2025. Retrieved from https://uksa.statisticsauthority.gov.uk/about-the-authority/strategy-and-business-plan/statistics-for-the-public-good/

United Nations Sustainable Development Group, UNSDG. 2019. United Nations Sustainable Development Cooperation Framework, June 2019. Retrieved from https://unsdg.un.org/resources/united-nations-sustainable-development-cooperation-framework-guidance

United Nations (UN), Department of Economic and Social Affairs. 2019. UN National Quality Assurance Frameworks Manual for Official Statistics. Retrieved from https://unstats.un.org/unsd/methodology/dataquality/references/1902216-UNNQAFManual-WEB.pdf

United States Office of Personnel Management. 2011. Paperwork Reduction Act, Version 2.0, April 2011. Retrieved from https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/paperwork-reduction-act-guide.pdf

West, Brady T., 2011. “Paradata in survey research.” *Survey Practice* 4(4).

Wilson, Laura and Emma Dickinson. 2021. *Respondent-Centred Surveys: Stop, Listen and then Design*. Thousand Oaks, USA: Sage Publications.

Yan, Ting, and Kristen Olson. 2013. “Analyzing paradata to investigate measurement error.” *Improving surveys with paradata: Analytic uses of process information.* 73-95.

Yan, T., & Williams, D. (2022). “Response Burden – Review and Conceptual Framework.” *Journal of Official Statistics*, 38(4), 939-961. https://doi.org/10.2478/jos-2022-0041

CHAPTER 2
NEEDS, SCOPE, AND BUSINESS CASE

Leesha Delatie-Budair (Statistical Institute of Jamaica)
Solly Molayi (Statistics South Africa)
Charles Lau (Gallup)

CHAPTER 2
NEEDS, SCOPE, AND BUSINESS CASE

2.1. Overview

1. Effective planning is the critical foundation for high-quality survey projects and programmes. It is essential that there is a clear and shared understanding of the survey's needs and expected outcomes, agreement on the scope, and adequate justification for proceeding. These steps are necessary for initiating new survey projects and programmes (see **Box 2.1**).

2. These steps are also crucial for existing surveys where survey managers should periodically consult, confirm the needs and scope, and update the business case. This ongoing process helps ensure relevance and clarity over time, which are key dimensions of quality in official statistics (United Nations, 2019) (see Chapter 1).

Box 2.1. Project versus Programme

A survey **project** is defined as a one-off survey, while a survey **programme** is defined as a recurring set of surveys conducted over time.

3. The requirement to continuously monitor needs, scope, and business case for existing surveys may be prompted by either:

[1] **Environmental changes**, such as budget cuts, new stakeholder needs (including information requests from external sources), new methodology, technology, or emerging issues, or

[2] **An information request**, which could also include an action plan from the evaluation of a previous iteration of the survey process, data needs for other statistical products, or best practices nationally, regionally, or internationally (UNECE Statistics Division, 2019). For example, government directives or new data requirements for national and global development frameworks sometimes require new surveys or changes to existing surveys.

4. This chapter outlines the process of initiating a survey prior to beginning work on implementing the survey. Once the survey's needs, scope, and business case are established and resources are allocated (this Chapter, then the process of survey design and implementation (covered in Chapters 3-10) can begin.

5. This chapter is motivated by the "Specify Needs" phase of the Generic Statistical Business Process Model (GSBPM) (see Section 2.1.4). The chapter covers foundational concepts when setting up surveys and serves as a critical precursor to the more technical chapters that follow. Particularly, this chapter introduces three key tasks National Statistical Offices (NSOs) undertake when planning a new survey:

- Needs: The required outcomes from the survey
- Scope: What the survey covers and the extent of survey activities
- Business Case: The argument in favour of conducting the survey

6. In this section, we briefly introduce each task (Sections 2.1.1 – 2.1.3). Then we cover each task in detail later in the chapter: Needs in Section 2.2, Scope in Section 2.3, and Business Case in Section 2.4. The primary audience for this chapter is NSOs, but elements of this chapter are also relevant for other types of organisations conducting surveys.

2.1.1. Needs

7. The first step when initiating a survey is understanding the needs and distinguishing them from wants. **Needs** are defined as "the differences between your current achievements and your desired

accomplishments" (Watkins, Meiers, & Visser, 2012). The need for a survey or survey module may be internally generated within the National Statistics Office (NSO), generated by a member of the National Statistics System (NSS), or other parts of the government (e.g., executive branch). Irrespective of the origin of the survey need, a thorough needs assessment should be conducted.

8. Conducting a needs assessment involves in-depth and systematic analyses aimed at identifying and defining gaps while charting a path for improvement. Identifying the needs will provide guidance during the development and implementation stages and serve as the guiding framework for operational activities related to the survey.

9. When specifying the needs, it is important to distinguish between **survey outcomes** (external needs) and **operational needs** (internal needs). For example, a survey outcome may be the production of statistics about vaccination coverage for a specific subgroup or trends in economic well-being: these outcomes are the result of the survey. Survey outcomes (external needs) are the needs of stakeholders. In contrast, operational needs are requirements that must be met (e.g., availability of funding or human resources) to achieve the outcome. Operational needs (internal needs) represent what is needed to conduct the survey.

10. There may be a need for a completely new survey, a need to add or modify a survey module[1], or a need to add or modify survey questions. These changes may be for a one-off survey (survey project) or a recurring survey (survey programme). Irrespective of the type of need, great care should be taken when conducting the assessment before proceeding with implementation.

11. NSOs are encouraged to establish prescribed procedures for conducting needs assessments to ensure that all factors are considered consistently during the survey initiation stage and are appropriately documented and addressed. **Figure 2.1** summarizes the benefits of a formal needs assessment process. Section 2.2 explores detailed strategies for conducting needs assessments.

Figure 2.1. Benefits of a formal needs assessment process. (Watkins, Meiers, & Visser, 2012)

2.1.2. Scope

12. Defining a survey's scope involves clearly articulating all the major decisions related to survey design - e.g., geographic scope, target population, conceptual framework, questionnaire length and content, languages, sampling design). Further, the scope also must clearly describe deliverables (scope of work) -

[1] A "survey module" is a set of related survey questions designed to measure a particular phenomenon.

e.g., whether data analysis is included (see Chapter 9), and if so, what type of outcomes and what dissemination activities (see Chapter 10) are included. Defining the scope in detail is important for establishing the resources (budget, time, human resources) required to implement the survey while successfully maintaining quality standards. Recommendations on defining and managing the scope will be discussed in greater detail in Section 2.3.

13. Clearly defining and documenting the scope of the survey project is essential. The scope is informed by the needs assessment and is often an iterative process until agreement is reached. Ensuring all stakeholders have a shared understanding of the scope is paramount to a successful outcome and for managing scope creep. Documenting the scope of work (that is, the specific activities to be undertaken), and the survey's scope (that is, what the survey will cover, helps manage expectations) ultimately leads to improved customer satisfaction. Whether the survey need is internally or externally generated, the success of the survey is dependent on ensuring that the survey outcomes (external needs) are aligned with stakeholder expectations. The scope of the survey should, therefore, be clearly documented to ensure expectations are aligned.

2.1.3. Business case

14. *"A business case is a value proposition for a proposed project that may include financial and nonfinancial benefits"* (PMI, 2021). A well-prepared business case is crucial for gaining approval and support for surveys, ensuring they are viable, strategically aligned, and well-planned. The business case should be clearly documented and shared with key stakeholders. Typically, the business case document contains information on the justification for the project and the expected value to be obtained from undertaking the survey.

15. The business case forms the basis for all other decisions regarding the survey, including approaches to design, integration, analysis, and dissemination. It helps identify the benefits of undertaking the survey, the resources required, and the alignment with strategic priorities. By monitoring alignment with the business case, the survey team can make informed decisions to meet or exceed the value proposition.

16. As the data ecosystem evolves, preparing and presenting robust value propositions for conducting household surveys that clearly identify the return on investment will become increasingly important. Over time, it has become increasingly difficult to execute surveys given rising costs, declining response rates, increased privacy concerns, rapidly evolving technology, diversity of data sources, respondent burden, and diminishing data quality. In this rapidly evolving landscape, the approach to initiating surveys must take a more strategic approach, understanding and articulating the business value and preparing a solid foundation for access to funding and the use of scarce resources.

17. As the survey project progresses, the business case may need to be updated or refined in response to changing circumstances, additional information, or new opportunities. Updating the business case should be done in collaboration with key stakeholders to promote consensus and buy-in. The business case will also inform requests for funding and other requisite approvals that survey managers need before proceeding. It provides the cadence necessary to ensure activities are aligned, objectives are consistent, and value for money is achieved. Guidance on developing a business case is contained in Section 2.4.

2.1.4. Conceptual framework

18. This Chapter draws on the conceptual framework of the Generic Statistical Business Process Model (GSBPM). While other frameworks may also provide useful guidance, this handbook will focus on the GSBPM.

19. The GSBPM is a framework that describes and defines the steps of business processes needed to produce official statistics. Adhering to phase 1 of the GSBPM is not just a step but a crucial foundation. This phase, "Specify needs," depicted in **Figure 2.2**, forms the basis for determining the necessity and feasibility of the survey. It involves specifying explicit requirements and delineating the operational processes necessary for generating official statistics. This process ensures the initial exploration and identification of the required statistical data, as well as the specific attributes expected from the survey.

20. It is important to consider this Handbook along with other important references (see Section 2.6. Resources) that offer valuable guidance and insights during the process of assessing data needs. Note that this chapter will not provide all the details of the GSBPM[2].

Figure 2.2. Specify needs phase, GSBPM 5.1.

2.1.5. Roadmap

21. This Chapter addresses three components of survey planning:

- **Needs (Section 2.2),** including recommendations and best practices for identifying data needs, including case studies highlighting these practices. Once the data needs are identified, the chapter emphasizes the importance of confirming and consulting, along with best practices for stakeholder engagement.
- **Scope (Section 2.3)**, including defining and managing the scope and presenting guidelines and best practices.
- **Business case (Section 2.4)** preparation and submission, including examples and best practices.

22. Finally, the Chapter addresses key measures for ensuring quality (Section 2.5) and resources (Section 2.6) for identifying data needs, determining and managing the scope, and preparing the business case.

2.2. Identifying needs

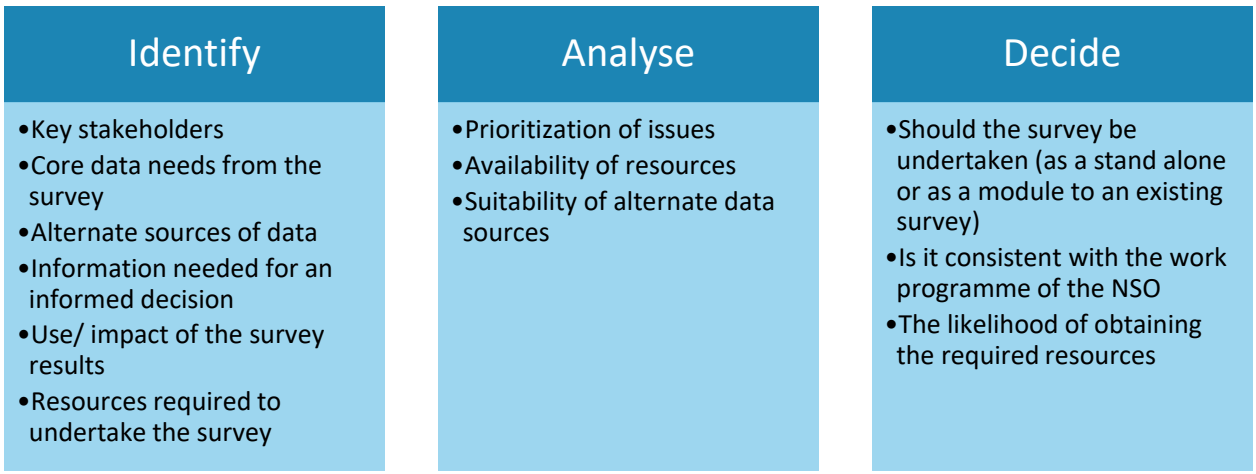
23. It is recommended that NSOs standardize their approach to conducting needs assessments. This will ensure that all aspects are consistently covered and serve as a guide for new survey managers. Generally, needs assessments include three broad sets of activities (see **Figure 2.3**):The "Identify" and "Analyse" activities are covered in Sections 2.2.1 and 2.2.2, and the "Decide" activity is covered in Section 2.2.3.

24. A needs assessment goes beyond simply asking stakeholders what they need. An effective needs assessment objectively identifies gaps and potential solutions. Identifying needs *"includes the initial investigation and identification of what statistics are needed and what is needed of the statistics"* (UNECE

[2] For more details on the GSBPM, please visit https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51

Statistics Division, 2019). Identify needs is the first sub-process of the GSBPM and may be initiated by a desire for additional information, whether internally or externally generated, or an environmental change.

Figure 2.3. Steps in conducting a needs assessment.



25. Identifying needs is an important first step in survey planning that sets a firm foundation on which the entire survey is built. The focus of the needs assessment should be on clearly documenting the data gaps and uses of the survey results (see **Box 2.2**). A good needs assessment will also serve to align the needs and expectations with available resources. It will also identify alternatives and provide adequate justification for undertaking the survey project or programme. This specification of the need(s) will inform the scope and form part of the business case.

Box 2.2. Tasks in a Needs Assessment
(Watkins, Meiers, & Visser, 2012)

A useful needs assessment will accomplish the following:

- Focus on results first, solutions second.
- Define needs as gaps in results.
- Align operational, tactical, and strategic performance.
- Systematically analyse needs to inform decisions.
- Consider a broad array of possible activities.
- Compare activities against performance criteria.
- Provide information that justifies the decision before it is made.

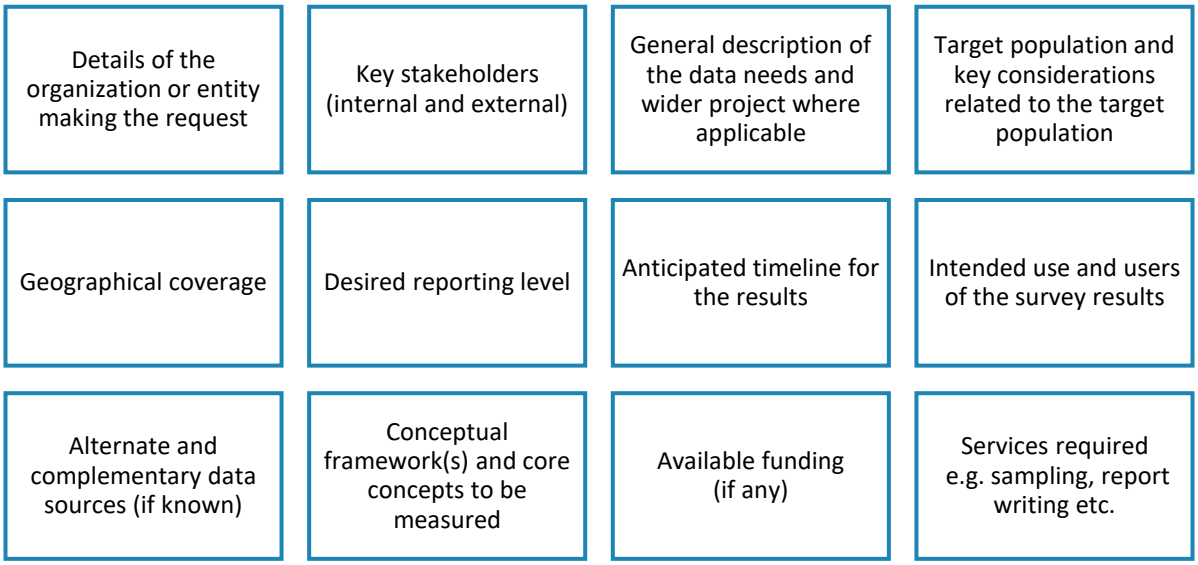
2.2.1. Identify and analyse: Survey outcomes (external needs)

26. When identifying needs, gaps between the current state of available data and the desired survey outcomes should be identified. This process involves prioritizing issues and solutions, along with systematically evaluating alternatives. In the context of surveys, a needs assessment includes reviewing alternate data sources and conducting a cost-benefit analysis of the impact of data collection against the resources required to execute a survey effectively. Essentially, a needs assessment distinguishes the truly necessary results from a survey, considers whether there are alternatives, and identifies the "nice-to-haves" versus the "must-haves."

27. The NSO plays a pivotal role in establishing a comprehensive statistical production framework which is aligned with the GSBPM. As part of assessing internal and external information needs, the NSO needs to establish a process for evaluating survey requests. This is effected through a streamlined process for

initiating requests for surveys. Countries are encouraged to develop systems for capturing these requests. This system could clearly document these requests, whether they originate within the NSO, from another government agency, or outside of government. **Figure 2.4** shows key elements to capture on the initial requests.

Figure 2.4. Desired elements of a request for a survey, survey module, or survey questions.



28. This information is best captured using a standardized form. The form should clearly articulate the extent of work the NSO may or may not undertake and the associated confidentiality restrictions. Reference should also be made to any relevant laws, regulations, or policies that govern the execution of surveys by the NSO. This process helps to streamline the process while also serving as a tool for the requesting entity to carefully consider their needs before making a request.

29. As stated earlier, a survey need could also emanate from an environmental change or other internal factors. It is recommended that even in instances of internally generated needs, the programme manager[3] takes the time to clearly document the emerging need by making a formal request to the section of the NSO in charge of surveys.

30. The initial request for a survey, survey module, or additional survey questions is a high-level request outlining the main goals and key outputs of the survey. It provides the general framework for future activities. The NSO should be informed of this during the ideation phase, which allows for more effective planning and implementation. Early dialogue with the NSO will improve the design and development of the survey and ensure the data needs are fully considered.

2.2.2. Identify and analyse: Operational needs (internal needs)

31. A survey needs assessment goes beyond external needs (survey outcomes) and also considers operational (internal) needs, which are requirements of the survey, such as resources, approaches, and organizational capacity. Needs assessment is a complex process involving several stakeholders and key considerations. It is a strategic initiative with tactical and operational considerations. The needs of a survey project or program include the operational needs, resource requirements, and data needs from the survey.

[3] The person in charge of the section within the NSO that requires the data.

The needs assessment often focuses on the survey outcomes of the survey. However, it is important to consider all needs, i.e., what data users wish to have and what your internal stakeholders need to deliver these outputs effectively.

32. Assessing the resource requirements will involve systematically analysing the operational, human, financial, and technical resources required to deliver the survey within the desired timeframe and scope effectively. The NSO must also determine whether the required resources are available in-house, whether additional support is required, or if the required resources may be procured in time for the survey. For example, an NSO seeking to execute a survey using CAPI must consider whether there are sufficient devices available within the NSO, if the NSO has the IT support staff to code the electronic questionnaire, and where these are not available if the procurement process will be concluded in time for testing the survey.

33. The results of a meticulous assessment of the NSO's capacity to undertake the survey will inform the preparation of the scope, survey budget, timelines, and business case.

2.2.3. Decide: Evaluating requests for surveys or survey modules

34. Having identified the need, the next step is to review and assess the request to determine whether the activity may be undertaken. This step may include a new survey or survey programme, a new survey module, or additional questions on an existing survey.

35. **Data availability.** Before proceeding with a survey, survey module or a set of survey questions, it is important to critically assess the data landscape to determine whether there are alternate data sources that address the identified need. It is important to assess the reporting level, timeliness, and periodicity of those alternate data sources, how accessible they are to the NSO, and whether data integration is possible. Oftentimes, stakeholders request modules because of a desire to cross-tabulate and co-analyse different phenomena. However, given advances in data science, it may be possible to integrate various data sources to yield more enriched analysis (see Chapter 1). The potential for data integration is a key consideration when assessing data availability.

36. **Reporting obligations and data requirements.** Key considerations when conducting a needs assessment include international, regional, and national priorities and frameworks, best practices, the mandate of the NSO, the NSO's capacity, and alignment with the country's development plan. This comprehensive approach will ensure the effective evaluation of survey needs aligned with overarching goals and strategies at various levels.

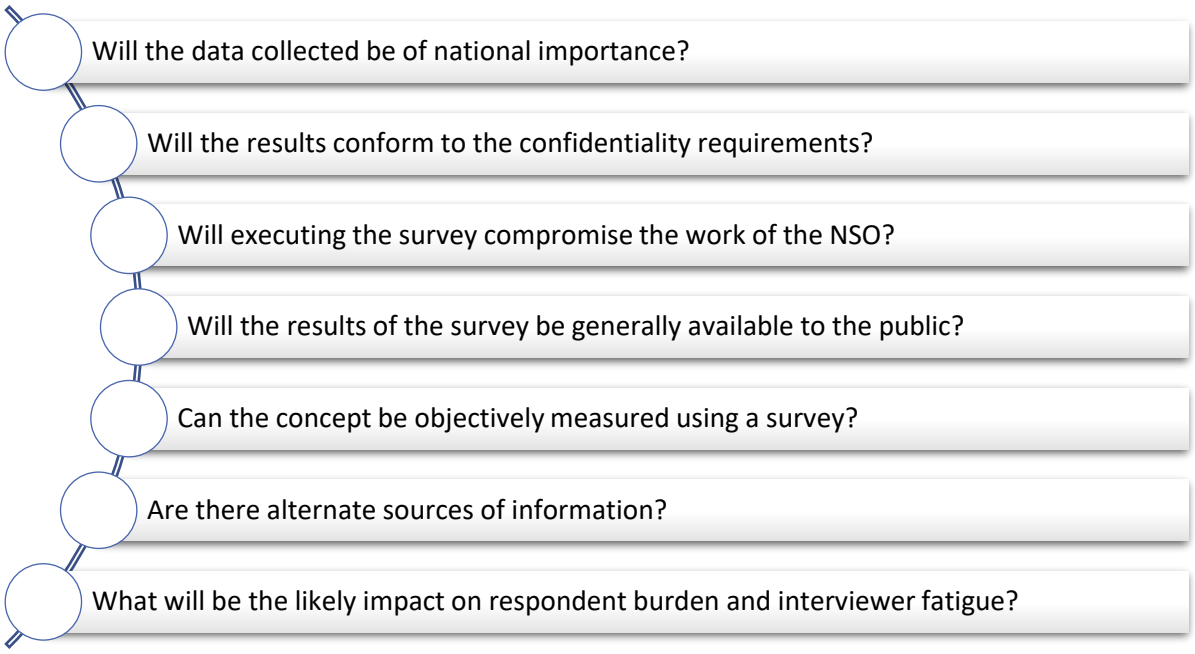
37. **Development frameworks**, such as the 2030 Agenda, serve as a proxy for the demand for statistically significant national information. Typically, specific departments or subject matter divisions within the statistical organization and other stakeholders in the NSS are responsible for identifying these needs.

38. **Other key considerations.** Once external and internal needs have been identified, great care should be taken in deciding whether a survey is the best tool to garner this data. Survey requests should also be considered holistically to ensure efforts are not duplicated. As best as possible, surveys should be combined to avoid duplication and redundancy while minimizing respondent burden and interviewer fatigue. Combining surveys ensures the optimal use of scarce resources for the conduct of population-based surveys. Key questions to be answered when deciding on whether to proceed with a survey are shown in **Figure 2.5**.

39. Additionally, given the ever-declining response rates for surveys, the length of the survey must be considered. Engaging in discussions with stakeholders about striking a balance between survey length and data quality is crucial. Lengthy questionnaires can result in respondent fatigue, leading to the provision of

inaccurate information (see Chapter 4 on questionnaire design). Therefore, it is imperative for NSOs to establish a framework for prioritization and content rotation to ensure that identified needs are based on international, regional, and national priorities and that respondent burden and interviewer fatigue are managed. This framework should be meticulously designed to identify the strategic measurement priorities for the country and to discern any data gaps in the data supply.

Figure 2.5. Key criteria for evaluating requests for surveys, survey modules, or survey questions.



40. Having completed the initial evaluation, there are three possible outcomes (See **Figure 2.6**).

2.2.3.1. Selected case studies

41. In the following pages is a case study from South Africa about conducting needs assessments (**Box 2.3**). [[Future versions of this chapter will incorporate other case studies.]]

2.2.4. Consult and confirm needs

42. After completing the 3-step process above (Identify➔Analyse➔Decide), and determining how to proceed, it is essential to consult key stakeholders and confirm that their needs are understood.

43. NSOs generate official statistical products used by a wide range of users, including government (national, provincial, and local), research institutions, professional bodies, media, businesses, educational institutions, and the general public. Once a desired survey outcome (need) has been identified, it is important to review whether other stakeholders have similar data needs. Population-based surveys are expensive to administer and very resource-intensive. Therefore, having received an initial request, it is important to consult with all stakeholders and confirm whether the decision reached about how to proceed will meet their needs. Additionally, consulting with the internal stakeholders of the NSO will allow for a better assessment of the resource requirements and organizational capacity to successfully execute the survey project or programme.

Figure 2.6. Potential outcomes of a needs assessment.

PROCEED:

- This means the NSO finds the initial request to be appropriate and is willing to move forward with the process.
- This may include further refining the data needs (confirming and consulting) as well as seeking the required resources (preparing the business case).

DEFER:

- Having received a request for a survey, survey module or survey questions, the NSO may conclude that the timing of the request may not be practical given available resources.
- In such instances, the NSO is encouraged to negotiate with the requisitioner to determine an appropriate timeline for implementation.

DECLINE:

- This means that the initial request is inconsistent with the scope of work of the NSO, or does not align with international, regional or national statistical priorities.
- NSOs may receive several requests for surveys, survey modules or survey questions given their expertise in this area. However, fulfilling these requests must align to the organizations mandate, the international, regional or national statistical priorities and must not compromise existing statistical work programmes.

44. Developing a stakeholder engagement strategy that will guide this process is important. For example, appropriate channels and means of communication should be used because needs change, and different users respond to different engagement methods. Additionally, it is good practice to convene consultation meetings where groups of stakeholders can discuss the needs identified and confirm whether there are appropriate alternatives.

45. The "Specify Needs" sub-process 1.1 of the GSBPM underscores the importance of engaging in comprehensive consultations with internal and external stakeholders to meticulously ascertain the statistical data requirements. This sub-process underscores the significance of comprehending user needs to ensure that the statistical entity comprehends not only the anticipated deliverables but also the timing, methodology, and, notably, the rationale. A thorough grasp of user needs is pivotal for generating pertinent official statistics, as stipulated in Principle 1 of The Fundamental Principles of Official Statistics, which advocates for relevance, impartiality, and equitable accessibility, aiming to provide users with high-quality information (UNECE, 2013).

Box 2.3. Needs Assessment Case Study from South Africa

Statistics South Africa developed an INTEGRATED INDICATOR FRAMEWORK (IIF) after reviewing various development frameworks:

[1] The National Development Plan (NDP)

[2] Sustainable Development Goals (SDG)

[3] African Union Agenda 2063

[4] Medium Term Strategic Framework (MTSF). (see FIGURE)

The IIF is the comprehensive outcome of an indicator framework review process that covers all the measurement requirements of each framework. It creates a single consolidated framework containing the national important measurement needs, and critically, rationalises duplicate indicators where possible.

Additionally, the IIF will organise indicators along with common targets and goals. In its final form, the IIF will represent the total demand for statistical information of national importance and hence generate a proxy estimate for the "Data Demand" variable.

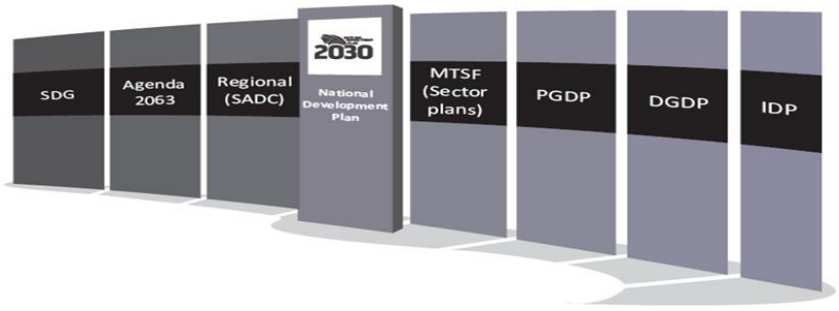
Since the IIF cuts across all statistical domains, its creation will require a system-wide response by the data producers in the NSS. The development of the IIF also provided Stats SA with the opportunity to map IIF indicators to data sources that were already identified, for example, in the SDGs. By so doing, the IIF presented Stats SA with an opportunity to compile a register of data sources that respond to the IIF and start the process for measuring the data supply in the NSS.

The objectives for the development of an IIF

- To reduce the burden of reporting
- To apply the same rigorous technical processes of the SDGs
- To help identify duplicate data sources within the system
- To mitigate the existence of contradictory estimates in the public domain

The framework not only enables the Statistician-General to coordinate the collection and dissemination of statistical information effectively but also establishes a structure for harmonizing the supply and demand of such information within the country. Moreover, it plays a crucial role in identifying and prioritizing the strategic areas for measurement within the country's statistical landscape. The IIF provides a thread across policy agendas.

FIGURE: NDP is at the Centre of IIF.



2.2.4.1. Confirm needs: Output objectives

46. The survey output should be agreed upon with stakeholders at a high level. The NSO is responsible for deciding on the methods based only on professional considerations (Principle 2 – UNFPOS). Having received the initial request, the NSO should, in consultation with key stakeholders, agree on and document the following:

[1] The policy/programme objectives from which the data need arises.

[2] Desired use of the survey output

[3] Unit of analysis

[4] Desired frequency

[5] Other desired survey output

2.2.4.2. Other considerations

47. Best practices when consulting and confirming data needs include the following:

- Examine best practices, methods or standards from regional and international partners that produce similar statistics
- Stakeholder consultations (internal and external), e.g. what to ask when trying to determine data needs (not asking for sample questions, but focusing on how the data will be used, etc)
- Prepare a stakeholder consultation report to ensure that statistical products meet user needs
- Prepare a needs assessment report for a technical review
- Ensuring the user knows what it is expected to deliver and when, how, and perhaps most importantly, why
- Ensuring the activity aligns with the organisation's mandate
- Fitting the survey project into a work programme
- Identifying and assessing alternate data sources, including assessing data gaps, introducing data integration and discussing real vs perceived gaps
- Integrate statistical and geospatial information, considering the advantages and disadvantages of different types of geography

2.2.4.3. Alignment on frameworks and standards

48. A critical part of the stakeholder consultation and confirmation process is to align on what conceptual frameworks and international standards will be applied to the survey. The work of the NSO should conform to international standards and best practices. Therefore, assessing whether guidelines already exist to measure the specified concept(s) is essential. Adaptation may be made for local circumstances, but as best as possible, the conceptual frameworks guiding the measurement of the phenomena should be stated and agreed upon at the onset of survey development. Conceptual frameworks will inform other technical tasks such as the questionnaire development (Chapter 4), data processing (Chapter 7), and analysis (Chapter 8) phases. See Part II of the Handbook for examples of conceptual frameworks and international standards for specific topics.

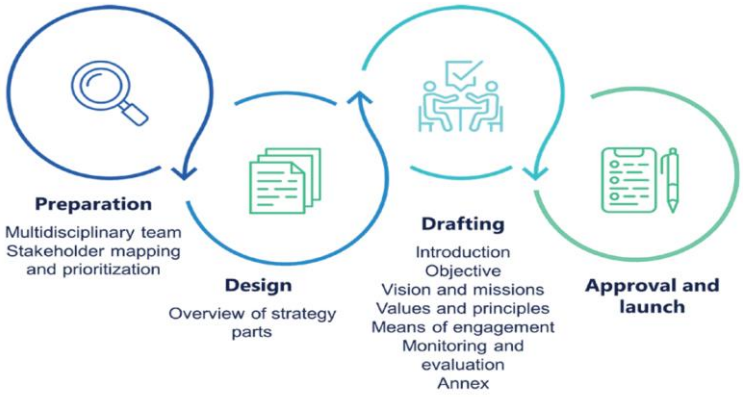
2.2.5. Selected case studies

49. This sub-section outlines the selected regional and country case studies with user engagement processes. **Box 2.4** summarises an integrated user engagement strategy from UNECA.

Box. 2.4. Integrated User Engagement Strategy, UNECA

In 2020, the United Nations Economic Commission for Africa developed and published guidelines for formulating an integrated user engagement strategy for national statistical systems. The primary objective of these guidelines is to aid national statistical systems in creating robust and strategic user engagement programs. The process of developing a user engagement strategy to facilitate the establishment and sustenance of high-quality dialogue and interactions with users can be delineated into four distinct phases: preparation, design, drafting, and approval and launch.

FIGURE: Process of developing a user engagement strategy



User engagement can and should be employed at every stage of statistical production in a continuous process of querying, feedback and response. With practice, user engagement will become integrated into the core business of the data value chain.

Activities for engaging users can be divided by their purpose into five groups: establishing contact with users; establishing user needs; consulting users; instigating user engagement; and establishing a feedback loop and building statistical literacy.

2.3. Determining the scope

50. The scope includes two broad components:

- the high-level survey design
- the statement of work

51. Determining the scope builds on the foundation of the needs assessment and represents the agreement on the way forward. Defining the scope of a survey is crucial for its success. It involves clarifying the purpose, defining the key stakeholders, and agreeing on the information to be gathered. The scope acts as a roadmap, guiding the survey's development and ensuring it aligns with its objectives. It builds on the needs assessment and represents an agreement on what the survey will achieve. Having understood what is needed, the scope of the survey speaks to what can be delivered within the agreed timeframe. Unfortunately,

these two may not align perfectly; however, a clearly defined scope is very useful in managing expectations and setting boundaries.

2.3.1. High-level survey design

52. The scope of a survey must include a high-level agreement on what the survey will look like. At a minimum, this should include an agreement on:

- What topics the survey will cover?
- Which conceptual framework(s) are to be used?
- Who will be surveyed (target population, geographic coverage etc.)?
- What type of sampling approach will be used?
- What mode will be used?
- When will the survey be fielded?

53. During the survey design phase, these elements will be further refined as outlined in Chapters 3-10. However, general agreement on key design decisions sets the parameters for future survey-related activities and also helps to prepare the business case. The governance framework and survey management strategy may also be agreed on at this stage.

2.3.2. Statement of work

54. In most instances, the NSO will execute a survey in its entirety. That is, the NSO will undertake all activities related to the survey described in Chapters 3-10 of this Handbook. However, there are many instances where the NSO will undertake a survey on behalf of another member of the National Statistical System (NSS). In these instances, the NSO may undertake some, but not all all activities related to the survey.

55. In cases where the NSO is conducting a survey on behalf of another entity, the scope should include a Statement of Work, which is a clear agreement on the activities to be undertaken by the NSO. There are several groups of activities that an NSO may undertake based on internal policies and best practices. For example, a NSO may conduct the sample design and selection only, and another entity in the NSS may perform the rest of survey activities. Other times, a NSO conducts all aspects of the survey, but another entity may conduct analysis and dissemination. The available sets of activities that the NSO is willing to undertake should be clearly defined and communicated to stakeholders.

56. Defining the Statement of Work is important, as it impacts the final output shared externally by the NSO. For example, the NSO may be engaged to execute data collection for a survey. However, given the confidentiality provisions, this may include minor data processing as survey respondents are de-identified. Activities may also include some questionnaire design and sampling, as the NSO must ensure that the data being collected is of the highest quality. It is important to clearly document and agree upon the complete set of activities that the NSO will undertake and the impact that this will have on the budget, the output that is produced and what is shared with stakeholders.

2.3.3. Managing the scope

57. Effective management of scope is essential regardless of whether requirements are documented upfront, evolve, or are discovered. New requirements may emerge during the survey planning process, and needs may be clarified. This may impact the scope of the survey. A critical part of managing the scope of a survey is identifying an individual responsible for effectively managing the survey to achieve success. This person may be referred to as the Survey Manager (see Chapter 3).

58. It is essential that the Survey Manager closely monitor any changes in the survey needs and design and carefully documents and communicates them. Poor requirement management can result in rework, scope creep[4], user dissatisfaction, budget overruns, and project failure. The scope, schedule, cost, resource needs, and risks should be well-defined early in the project life cycle and remain relatively stable.

59. As new requirements emerge, it is crucial to develop estimates of the impact of these requirements on the budget and timeline, as well as any additional resources that may be required. This must be communicated clearly to the client (project sponsor) as well as internal stakeholders. A formal decision should be taken regarding whether to proceed with the changes or to maintain the original scope. A well-defined work breakdown structure and a comprehensive scope management plan (see Chapter 3) are vital components for defining, developing, monitoring, controlling, and validating the project scope (Project Management Institute, 2021).

60. It is recommended that periodic meetings with internal and external stakeholders be held throughout the survey lifecycle. However, a review after each major activity (milestone) is essential. This is critical for managing scope creep and setting and adjusting expectations.

2.4. Preparing the business case

61. The Business Case is a document outlining the arguments in favour of conducting the survey or adding the survey module or questions. It is the culmination of the activities outlined in the Needs and Scope sub-section, providing documentation of the information gathered so far in a manner geared towards obtaining the requisite approvals. The Business Case outlines key considerations and resource requirements, including the budget, organizational capacity, as well as the implications of not doing the survey. Depending on the approval process at the country level, the business case may be an internal document, as well as it may be required to prepare a business case for submission to the Finance Ministry, Funding Agency or Development Partner. Irrespective of the recipient, the business case must provide a solid and coherent argument to conduct the survey or add the module. The primary purpose of the business case is to secure the requisite approvals and funding for the survey or survey module.

2.4.1. Typical elements of a business case

62. There are three key elements of a business case (PMI, 2021). These elements are adapted in **Figure 2.7** for the context or surveys.

63. The business case is more comprehensive for a new survey project or programme. However, for repeated surveys, the business case may be updated iteratively based on the evaluation results of previous survey rounds or environmental changes. The Survey Manager should review and update the Business Case as needed.

64. For an example of what content is included in a business case, see a case study from South Africa in **Box 2.5**.

[4] Scope creep is defined as incremental changes in the scope of a project, including changes in the design, output, work to be undertaken, timelines etc.

Figure 2.7. Key elements of a business case.

Business Need	Contains the rationale for conducting the survey or adding the survey module. It outlines the goals and objectives, as well as the status quo.
Project Justification	This explains why the survey or survey module is worth the investment, and why it should be done in the specified timeframe. It typically also includes a cost benefit analysis, and key underlying assumptions. It may also contain an analysis of the implications of not doing or delaying the survey or survey module.
Business Strategy	This contains the roadmap for achieving the desired outcome, and a statment of the key deliverables. The Business Strategy section of the Business Case contains an assessment of the organizational capacity, governance framework and key partners for the successful execution of the survey or survey modeule.

Box 2.5. Required Elements for a Survey Business Case in South Africa

Statistics South Africa has mapped the business process of compiling and approving business cases. The following should be included in the business case:

[1] Background
[2] Purpose and Scope
[3] Key Objectives
[4] Stakeholder Needs Analysis
[5] Cost
[6] Benefits and Outcomes
[7] Assumptions
[8] Constraints
[9] Risks and Mitigation Plan
[10] Timeframe
[11] Resources – Technical and Human Resources

2.5. Guidelines for ensuring quality

65. This section outlines the quality guidelines for the "Specify Needs" phase of the GSBPM as per Statistics Canada's Quality Guidelines Sixth Edition (Statistics Canada, 2019).

66. **All statistical processes:**

- Identify and analyse the information needs of internal and external users in relation to new information requests or the required environmental changes.
- Compare similar statistical operations, particularly standards and methods, used in other regional and international statistical organizations.
- Identify the needs of specific user groups, such as people with disabilities or ethnic groups. • Examine information needs for the best short and long-term solutions.
- Consult users systematically and extensively to clarify content and generate support from project partners.
- Establish and maintain relationships with data users in all sectors to improve information relevance as well as the dissemination of products and services.
- Identify and define operational constraints, such as the reference period, costs, resources and data collection methods.
- Establish production goals together with users and key stakeholders.
- Include measurable quality dimensions in the statement of objectives.
- Consider the objectives and needs of subsequent or parallel statistical activities when determining production objectives.
- Develop measurable concepts according to the statistical information to be produced (target population, statistical unit, etc.). The concepts do not need to be aligned with existing statistical standards at this phase.
- Analyse available and accessible sector data in terms of relevance, frequency, quality, timeliness, etc.
- Determine whether record linkage is a viable option by starting to identify which datasets could be linked.
- Assess the legal framework of all possible sources concerning collection and use.

2.5.1. Quality indicators for the Generic Statistical Business Process Model (GSBPM)

67. A fundamental role in quality management is played by a set of quality indicators that should be implemented within the sub-processes to prevent and monitor errors. The main goal of quality management within the statistical business process is to understand and manage the quality of the statistical products. The Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources document outlines a set of quality indicators that have been developed for the production of statistics from both survey and administrative data sources (ADS), regarding the different stages of the Generic Statistical Process Model (GSBPM) Version 5.0. (UNECE, 2017, UNECE 2013).

2.5.2. Quality dimension and indicators at this phase

68. The quality dimension and indicators at this phase are summarized in **Table 2.1**. Considering the "relevance" dimension, for example, quality indicators may include:

- A description of user needs and how users intend to use the data
- Analysis plans that include a description of tables to be released
- A business case establishing the gap between user needs and intended outcomes.

Table 2.1. Quality dimension and indicators at this phase.

Quality Dimension	Indicator	Notes
Adequacy of resources	To what extent have resource requirements for the proposed outputs and their quality measures been considered?	Includes extreme value checks, population unit checks, variable checks, combinations of variables checks, etc.
Adequacy of resources	Has the data source been evaluated in terms of its cost-effectiveness?	
Relevance	To what extent does the business case conform to the requirements of the approval body?	
Relevance	To what extent does the business case reflect the findings, recommendations and proposals from steps 1.1 to 1.5 of GSBPM?	

2.6. Resources

[[To be added in spring 2025]]

2.7. References

Project Management Institute (PMI). (2021). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*. 7th edition. Newtown Square, PA: PMI.
The PMBOK Guide offers frameworks for defining and managing project scope, which can be adapted to survey projects. It includes techniques like work breakdown structures and scope statements.
https://ibimone.com/PMBOK%207th%20Edition%20(iBIMOne.com).pdf

Statistics Canada. (2019). Statistics Canada Quality Guidelines sixth edition. Catalogue no. 12-539-X. Ottawa, Ontario.
https://www150.statcan.gc.ca/n1/en/pub/12-539-x/12-539-x2019001-eng.pdf?st=NJEGtOft

United Nations. (2008). Designing Household Survey Samples: Practical Guidelines, Studies in Methods (Ser. F), No. 98. New York: United Nations.
https://doi.org/10.18356/f7348051-en.

United Nations. (1985). National Household Survey Capability Programme. New York: United Nations. https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Handbooks/surveys/National_Houshold_Survey_Capability_Programme-E.pdf

United Nations Economic Commission for Africa (UNECA). (2020). Guidelines for Developing an Integrated User Engagement Strategy For National Statistical Systems. Addis Ababa, Ethiopia. https://archive.uneca.org/sites/default/files/PublicationFiles/guidelines_for_developing_an_integrated_user_engagement_strategy.pdf

United Nations Economic Commission for Europe (UNECE). (2013). Generic Statistical Business Process Model. New York: United Nations. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

United Nations Economic Commission for Europe (UNECE). (2009). Statistical Metadata in a Corporate Context: A guide for managers. New York: United Nations. https://unece.org/info/publications/pub/21910

United Nations Economic Commission for Europe (UNECE). (2013). Generic Statistical Information Model (GSIM): Specification. New York: United Nations. https://unece.org/sites/default/files/2023-11/GSIM%20v1.2%20-%20UML%20Diagrams_1.pdf

United Nations Economic Commission for Europe (UNECE). (2017). Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources. New York: United Nations. https://statswiki.unece.org/download/attachments/185794796/Quality%20Indicators%20for%20the%20GSBPM%20-%20For%20Statistics%20derived%20from%20Surveys%20and%20Administrative%20Data%20Sources_Final.pdf?api=v2

This page has been intentionally left blank.

CHAPTER 3
SURVEY MANAGEMENT

Annotated Outline

The full Chapter will be written in Spring 2025.

CHAPTER 3
SURVEY MANAGEMENT

Annotated Outline

1. This Chapter reviews principles and practices of managing a survey. This chapter explores the essential aspects of survey management, defined as the oversight, direction, and coordination of survey projects. It covers key functions like planning, organizing, leading, and controlling while emphasizing the central role of managing people in this labour-intensive and multidisciplinary process. Survey management bridges planning and technical implementation, ensuring tasks are completed on time, within budget, and to specification. The chapter also highlights how management varies based on organizational context, project type, and other factors, and it defines success through effective stakeholder engagement and balancing time, cost, and quality. Topics such as scheduling, budgeting, risk management, and quality assurance are addressed alongside cross-cutting themes like ethical considerations and stakeholder management, offering a comprehensive guide for successful survey execution.

An annotated outline is presented here. The full Chapter will be written in Spring 2025.

3.1. Introduction

3.1.1. Definition and key functions of survey management

2. Survey management is defined in this section of the chapter. The introductory paragraphs present the key stages or functions of survey management.

3. **Definition:** Survey management refers to the activities conducted to allow a survey to meet its objectives. These activities are those in the process of monitoring the correct design, implementation, administration, and supervision of the analysis of data collected. Managing a survey involves several stages, from planning and creating the survey to distributing it, monitoring responses, delivering data for processing and analysis, and closing the project. Survey management also refers to the correct management of resources, particularly budget, schedule and staff. Ultimately, good survey management is about ensuring that the entire process is well-organised, efficient, and effective in collecting high-quality data that meets the survey's objectives.

4. **Challenges: S**ome challenges for successful management of surveys will be presented.

5. **Variation:** This section will present the differences between survey projects. While some projects have similarities, management can look different for different NSOs and/or organizations in charge of the survey. Many of the differences between surveys result from distinct organisational context such as the presence of governing boards or steering committees and higher management. At the same time, while large organisations usually have specialized departments in charge of different stages of the survey (design, sampling, field operations, etc), smaller organisations might have a team in charge of all stages of the survey.

6. Further, this section also highlights the challenges and differences between a survey and a survey programme. Survey management is after all, about managing a team or teams of people in charge of different stages of the survey. This section also includes the perspective when external providers of services are used for some tasks of the survey.

3.1.2. Success in survey management

7. This part of the section refers to basic indicators of successful survey management. In particular, the relevance of support the completion of the project in time, within the budget and with quality results. Along the process, it is important to maintain stakeholders engage in the process present reports and request approval of possible changes.

3.1.3. Survey management and other chapters in the Handbook

8. This part of the chapter discusses the interactions between Survey Management and other stages of the survey. The first key distinction should be that survey management comes after the planning process for the survey, where the needs are assessed, the scope is defined, and the business case is made (Chapter 2). Once the survey is approved and the work begins, the Survey Management task (this Chapter 3) begins. In practice, there may be overlap between Chapters 2 and 3: Some aspects of survey management may begin during the planning process. However, for the purposes of organizing this Handbook, this chapter divides Chapters 2 and 3 in this way.

9. Further, this section will also distinguish Chapter 3 from Chapters 4-10. Survey Management is not about completing the technical aspects of the survey (discussed through chapters 4-10) but making sure those aspects or tasks are completed.

3.1.4. Chapter roadmap

10. This section will provide the reader with an overview of the chapter, including a list of the various sections in the chapter. These include:

- Governance (Section 3.2)
- The Survey Manager (Section 3.3)
- Survey planning (Section 3.4)
- Technology (Section 3.5)
- Executing, monitoring, and controlling the project (Section 3.6)
- Survey project closing (Section 3.7)

3.2. Governance

11. Surveys are complex projects, but depending on the organization conducting the survey, different committees, boards, management and other bodies are important for the project. Depending on the size of the NSO or organization, different boards or committees could be in place. These bodies oversee a variety of factors such as ethical approval, regional and field operations, internal audits and some other elements. Activities associated with organizational context include stakeholder management and communication throughout project life.

3.3. The Survey Manager

12. This section defines the role and desired characteristics of a Survey Manager and the survey team. A survey team is composed by a group of specialists on questionnaire design, statics, IT staff, fields operations, and other areas (communications, etc.). The role of the survey management is to coordinate all members of the team to complete their assigned tasks on time and with the resources allocated to them. The survey manager if the person responsible for the survey and is in constant communication with the board members (stakeholders).

13. This section will outline the specific responsibilities of the Survey Manager and also highlight some attributions or qualifications of Survey Managers.

3.4. Survey planning

14. This section describes the process of planning a survey, including (1) project initiation, (2) overall planning, (3) budgeting and scheduling, (4) risk assessment, (5) ethical and other approvals, and (6) human resource planning.

15. The Survey Manager coordinates the survey team to elaborate a plan to complete the survey. In this process, the manager must identify all stages of the survey and their dependencies, the required resources and the schedule for each task. While the implementation of a survey usually relies on a team, the type of management will determine the specific tasks, dependencies and responsibilities of each member. The planning stage also covers the identification of risks associated with the different stages of the survey project and the development of a contingency plan. Planning of the survey also implies planning for quality and quality monitoring. At the planning stage, the key performance indicators should be defined for each stage of the survey.

3.4.1. Project initiation

16. From Needs, Scope, and Business Case (Chapter 2) to the implementation of a Survey: The first step in this process is for the Survey Manager to take the Business Case and transform it into a project. To do so, the Survey Manager should identify the objectives, main variables of interest, deadlines and total budget available. In this stage is that stakeholders are identified.

17. Different types of management: Based on the organisational context, survey management can take different forms: from specialized teams in charge of specific tasks, to a survey team engaged in all stages of the process. The survey manager should identify these characteristics and design a management plan that aligns with the organisation's structure and resources.

3.4.2. Overall planning

- Planning for dependencies
- Planning for changes
- Planning for quality
- Planning for tasks and roles (that is, defining the Work Breakdown Structure)

3.4.3. Budgeting and scheduling

- Relevance of the budget: Resources and staff needed
- Steps in building a project schedule
    - Process of estimating task duration: how to calculate quantitative duration; how to define dependencies among tasks; what are the constraints (imposed dates, key events/major milestones/resource availability)
- Tools: Schedule networks/Critical Path Method (CPM); GANTT chart
- Key considerations: Balancing resource and scheduling; reducing project completion time

3.4.4. Risk assessment

- Risk matrix

- Contingency and risk mitigation plans

3.4.5. Ethical and other approvals

- Ethical approvals
- Legal approvals
- Financial approvals

3.4.6. Human resources

18. This section covers the activities related to the identification of required roles and responsibilities for the staff, negotiation with the Human Resources department (for some organizations) or the company providing specific services (if external suppliers are used), conflict resolution, and general personnel management approaches.

3.5. Technology

19. This section will discuss the use of technology in a survey project and survey management. All stages of the survey require specific technological solutions, and the survey manager should take that into account for planning, budgeting and scheduling. At the same time, technology itself can be used for managing the survey, depending on the budget and adoption of technology by the organization, these can be achieved by using simple tools such as spreadsheets or more advanced software.

3.6. Executing, monitoring, and controlling the project

20. This section focuses on the ongoing management process of the survey project or survey programme. Here, the discussion on some tools and systems to monitor progress and references to some indicators (covered in Chapters 5-7) are discussed. Also, this section presents recommendations to manage change, as well as assess and apply changes in the project, budget, staff rotation and contingencies.

3.7. Survey project closing

21. Project closing refers to all activities the Survey Manager has to undertake once all stages of the survey have been completed. These activities include administrative closure: final budget and balance, staff liberation/termination, archiving; and submitting the final report to the stakeholders or clients.

This page has been intentionally left blank.

CHAPTER 4
QUESTIONNAIRE DESIGN, TRANSLATION, AND INSTRUMENTATION

Julie de Jong (ICF)
Luis Sevilla (NORC at the University of Chicago)
Brian Kirchhoff (NORC at the University of Chicago)
Brita Dorer (GESIS-Leibniz Institute for the Social Sciences)
Peter Takyi Peprah (Ghana Statistical Service)

CHAPTER 4
QUESTIONNAIRE DESIGN, TRANSLATION, AND INSTRUMENTATION

4.1. Overview

1. The survey questionnaire is the integral tool used to collect data from respondents. In a survey questionnaire, individual survey questions are generally arranged and answered in a predetermined order, and the questionnaire may include both survey questions for respondents, as well as instructions for interviewers to gather additional observational data during the interview process.[1] The questionnaire is almost always programmed in an electronic data capture system for CAPI, CATI, and web surveys, but PAPI is still used in some contexts. The questionnaire is often transformed into a "survey specification", providing item-by-item instructions to facilitate programming into the data capture system, resulting in what is often referred to as the "survey instrument", the tool ultimately used to conduct the interview. Successful respondent engagement with the content of the questionnaire and the resultant data quality depends on the respondent understanding the questions, recalling relevant information, and providing accurate and honest answers. However, data quality also depends heavily on the quality of the questionnaire itself, as well as how effectively it is transformed into a survey instrument.

Box 4.1. What is included in this chapter?

This chapter covers the development of the entire survey instrument, including defining the research objectives, how to write survey questions, and the translation of the questionnaire if multiple languages will be used. It also covers the instrumentation of the questionnaire when electronic data capture will be used to administer the survey and provides guidance on software platform elements needed to collect high-quality survey data.

2. The definition of a quality questionnaire has evolved over time. For much of the 20th century, simply writing questions for immediate administration in a survey was adequate. In the past several decades, quality standards have risen due to a voluminous literature on the scientific basis of questionnaire design (Fowler 1995; Schaeffer and Dykema 2011; Ornstein, 2013). In the past several decades, there has been a shift toward understanding the cognitive processes involved in responding to surveys, the importance of aligning research objectives to corresponding survey items, and the role the social context plays in questionnaire design. Accompanying this transformation are significant advances in methods used in the review, translation, technical design, and testing of questionnaires, including incorporation of machine learning and web probing (Buskirk & Kirchner, 2020; Willis, 2019). Ultimately, the objective of the questionnaire design process may be condensed into the development of an instrument that collects high-quality data that are aligned with the research objectives while simultaneously not overly burdensome for respondents or interviewers.

3. As discussed in Chapter 1, Total Survey Error (TSE) provides a framework for how measurement error and representation error impact data quality at each stage of the survey life cycle, including the questionnaire design process. Measurement error due to questionnaire design occurs when the information collected from the questionnaire differs from the true values for the variables of interest; for example, due

[1] Although this chapter is focused on questionnaires for household surveys, the principles are also applicable for questionnaires used to gather supplemental data from other sources, such as data at the community level (e.g., with a community leader) or institutional level (e.g., from a health clinic, school, etc.).

to ambiguity in the survey question or answer categories or due to misunderstanding of specific terms included in the survey question. Representation error occurs in questionnaire design in three ways: 1) item non-response bias introduced by features of the questionnaire itself, such as questions that are difficult to understand or sensitive for respondents to answer; 2) questionnaire non-response bias resulting from breakoffs due to the questionnaire; and 3) errors in the development of the survey instrument, such as errors in the programming of response codes or skip logic (Groves et al., 2009). A good questionnaire is one in which such survey error has been reduced through application of questionnaire design best practices, including those discussed in this chapter.

4.1.1. Conceptual frameworks for questionnaire design

4. This section introduces three conceptual frameworks that help questionnaire designers identify and reduce sources of survey error. These conceptual frameworks build upon each other and are complementary – each with their own focus.

4.1.1.1. Cognitive response process

5. Tourangeau's four-stage model of the cognitive response process is a respondent-focused framework used to understand how respondents answer survey questions (Tourangeau & Rasinski, 1988; Tourangeau et al., 2000). It breaks down the process into four stages: comprehension, retrieval, judgment, and response. In the development phase, the cognitive response process provides questionnaire designers with a framework to create questions that are clear and easy to answer. In the testing phase, the framework helps to identify potential issues with question clarity, memory recall, judgment formation, and response selection, allowing for refinements before the survey is finalized, as discussed in Section 4.6 of this chapter. **Table 4.1** provides a description for each stage, along with an example for each as applied to a survey question about access to clean water.

Table 4.1. Cognitive response process for the example survey question: *In the past week, did you have access to clean water daily? (Yes / No)*

Stage of the response process	Description of the stage	Example application of the response process stage
Comprehension	Respondents interpret the meaning of the question to understand what is being asked	Understanding how "clean water" is defined
Retrieval	Respondents recall relevant information from memory	Remembering how often they had clean water available in the past week
Judgment	Respondents evaluate the retrieved information to form an answer	Deciding if their water is "clean" and whether their access to clean water is frequent enough to count as "daily"
Response	Respondents map their answer to the response options provided by the survey	Choosing "yes" or "no" based on their judgment

6. While the model remains a foundational framework in survey methodology, there have been several advances incorporated since its development. There is a recognition that the four-stage response process

represents an idealized framework that may not always align with the actual survey response process. In practice, respondents' cognitive processes can be far more complex and less linear than the model suggests. Indeed, satisficing theory posits that respondents may not always engage in thorough cognitive processing but instead provide satisfactory answers requiring minimal cognitive effort and potentially skipping stages of Tourangeau's model altogether, leading to shortcuts and potential biases in survey responses and illuminating motivational and social sources of error (Tourangeau, 2018; Dillman, 2019). Additionally, the model has been extended and complemented by additional and alternative theories. For example, Willis et al. (1991) proposed a fifth element to the four-stage process, **logical/structural errors**, which speaks to the vulnerabilities presented by misalignment between the question as written, and the research objective that it is intended to achieve. Consider the example question: *"Do your children attend a public or a private school?"* The question assumes that the household has children **and** that those children attend school, neither of which may be true for a household.

4.1.1.2. Questionnaire design, development, evaluation, and testing (QDET)

7. The Questionnaire Design, Development, Evaluation, and Testing (QDET) framework is researcher-focused and provides a structured approach to minimizing survey error in the questionnaire design process. The QDET process involves several key steps:

- **Questionnaire design**, where questions are crafted to be clear and relevant;
- **Development**, which includes refining questions through expert reviews, translatability assessment, and pre-testing;
- **Evaluation**, involving cognitive interviewing and pretesting to identify potential issues; and
- **Testing**, where the survey is field-tested to ensure it performs well in real-world conditions (Willis, 2019).

8. The QDET framework emphasizes the iterative nature of questionnaire development. Researchers are encouraged to continuously refine and adapt their survey instruments based on feedback and testing results. This iterative process helps to identify and correct ambiguities, biases, and other issues that may affect the quality of the data collected. By incorporating diverse evaluation methods, such as cognitive interviews, translatability assessment, and usability testing, along with more recent innovations such as web probing and the use of social media, QDET ensures that questionnaires are not only theoretically sound but also practical and user-friendly, providing opportunities for improvement of the questionnaire and survey instrument ahead of fieldwork, as discussed in Section 4.6.

4.1.1.3. Respondent-centred design

9. Like the cognitive response process, the respondent-centred design framework is also respondent-focused, but arguably more holistic in its approach. As outlined by Wilson and Dickinson (2022), the framework provides a mechanism to reduce respondent burden by prioritizing the needs and experiences of the respondents throughout the survey design process. While the approach draws on the foundation laid by Tourangeau and expanded by Willis and colleagues (Beatty et al., 2019) and also centred around the respondent, Wilson and Dickenson's method focuses on the respondent interacts with the survey instrument at each stage of engagement. Importantly, and in contrast to traditional questionnaire development, the starting point in respondent-centred design is the "discovery phase", which occurs before writing survey questions and involves both a better understanding of data users' needs as well as actively listening to respondents and understanding their mental models and experiences. By doing so, survey designers can create questions that are clearer, more relevant, and easier to answer, reducing the cognitive load and

making the survey experience more engaging and less taxing while collecting data that are salient for data users. Additionally, attention to instrument design to increase usability and integration of language inclusive of a broader set of identities ensures that surveys are accessible to a wider audience, further reducing burden while potentially also reducing the likelihood of item-level nonresponse bias.

10. Elements of these frameworks are critical in the overall questionnaire design process and culminate in several guiding principles in the field of questionnaire design that are critical for minimizing potential errors.[2]

Box 4.2. What makes a good questionnaire?

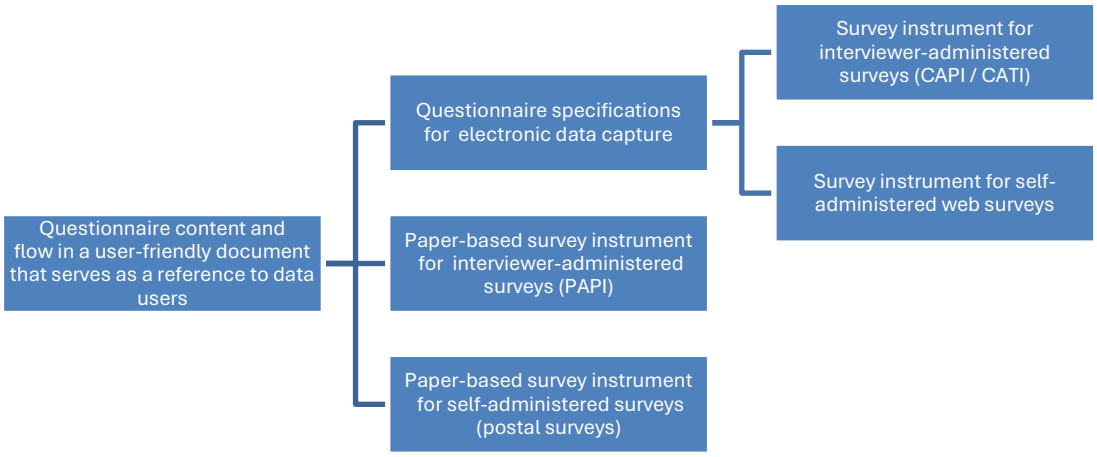
- The question items are designed with the target population in mind
- The questionnaire items accurately measure the constructs of interest
- All survey questions, response categories, and interviewer instructions in the survey questionnaire are reflected in the survey instrument
- The questionnaire is designed to minimize respondent burden by reducing the length where possible and using a logical and intuitive structure
- The instrument skip logic is programmed correctly
- Trend data are collected using the same wording over time
- All translations are equivalent to the source language
- Respondents understand the questions
- Respondents provide honest responses that are not impacted by social desirability bias
- Respondents are able to provide accurate responses to survey questions
- Respondents are able to answer all questions they are asked, resulting in low item non-response

4.1.2. Survey questionnaire development

11. The survey questionnaire is the document that includes the survey questions and response options. However, the specific steps required to produce the questionnaire in the format that will be used to collect data from respondents will vary, to some extent, by survey mode. In this chapter, the primary focus is on development of a questionnaire that is programmed into an electronic data capture platform and administered by an interviewer in a face-to-face survey (CAPI) or phone interview (CATI), using a laptop, tablet, or other electronic device, although there is also guidance provided for the technical development of questionnaires to be administered on paper. **Figure 4.1** illustrates the relationship between survey mode and the format of materials required for development.

[2] While a comprehensive review of approaches to ensuring quality in questionnaire design is outside the scope of this chapter, see Couper, 2008; Dillman et al., 2014; Fowler, 1995, 2014; Holbrook et al., 2003; Krosnick, 2018; and Tourangeau & Yan, 2007 for further resources.

Figure 4.1. Survey mode in the questionnaire development process.



4.1.3. Roadmap to chapter

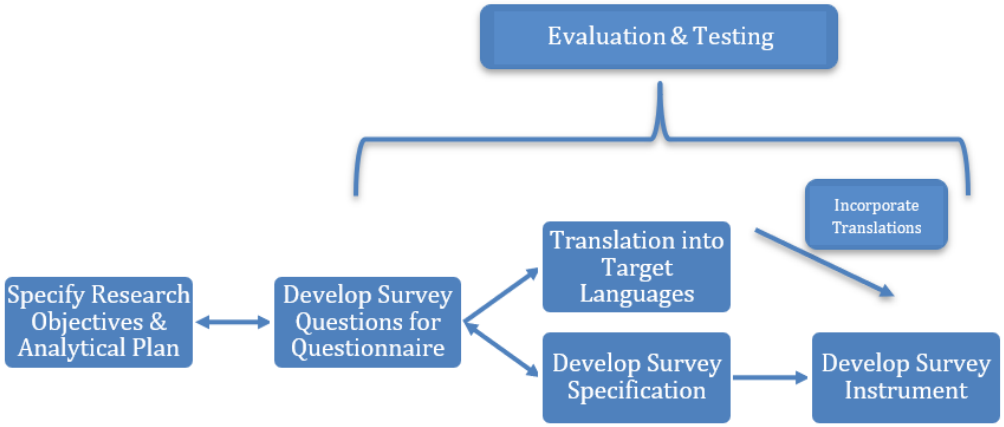
12. The questionnaire design process is not strictly linear; indeed, the testing and review required in a rigorous design process necessitate stage iteration. However, there is a general flow to the tasks in the development of the questionnaire and subsequent survey specification and survey instrument, which we follow in this chapter.

13. While the questionnaire development process is iterative, there are certain dependencies within and across tasks. As discussed in further detail in Section 4.2, it is critical to begin the design process by defining the research objectives and anticipated analysis so that survey questions result in data that address these objectives and are appropriate for the proposed analysis. Research objectives should be determined before beginning questionnaire design and continually referenced against the proposed survey questions throughout the questionnaire development process to ensure that all questions map to defined research questions, and that all research objectives, in turn, map to corresponding survey questions. This process may require refinement of research objectives as well; it may be that quality data to answer the initial research question cannot be obtained through survey research and the aims of the research require modification in order to design appropriate survey questions.

14. Depending on the survey target population, translation of the questionnaire from the source language – that is, the language in which the questionnaire is developed – into one or more target languages will also be required. If translation is required, survey questions must also first be developed in what is called the "source language" before translation can begin, although as discussed in Section 4.5, both translatability assessments as well as the early stages of translation may illuminate issues in specific survey items, requiring a return to the questionnaire for revision before the larger translation process begins. The source language questionnaire should be shared with stakeholders or key study personnel who are native speakers in the "target language(s)" (those languages into which the source questionnaire will be translated) to provide an opportunity to note any concerns anticipated with translation of key concepts. Similarly, a draft questionnaire is required for survey specification development, but issues uncovered in the development of the latter may necessitate a return to the questionnaire development stage for revisions. Evaluation and testing activities can and should occur throughout the entire development process, further requiring adjustments to the questionnaire, translation, survey specification, and survey instrument. **Figure 4.2**

illustrates the general process of questionnaire design through system instrument development, acknowledging the iterative process.

Figure 4.2. Dependencies within questionnaire development.



15. This chapter consists of the following sub-sections:

- **Section 4.2:** An overview of approaches to selecting and/or developing survey questions
- **Section 4.3:** Guidelines for technical design of the questionnaire when administered on paper
- **Section 4.4:** Guidelines for development of the survey specifications used to develop the electronic data capture system for the survey instrument when administered in non-paper format
- **Section 4.5:** An overview of the translation process and the available tools to produce and assess translations
- **Section 4.6:** An overview of the various approaches to questionnaire evaluation and testing, including expert review, cognitive pretesting, and survey pretesting
- **Section 4.7:** An overview of the components to consider when determining an appropriate data collection system
- **Section 4.8:** Best practices for programming and testing
- **Section 4.9:** Emerging approaches in questionnaire design, translation, and systems instrumentation at the time of the publication of this Handbook
- **Section 4.10:** A list of several electronic data capture software systems more commonly used at the time of the publication of this Handbook

16. Sections 4.2 – 4.6 provide an overview of the most critical elements in questionnaire development. As discussed in Section 4.2, the first step in questionnaire development is to determine the questionnaire content and flow, producing a document in a user-friendly format that can serve as a reference to data users. The next step will be determined by the survey administration mode. Section 4.3 provides guidance for surveys using paper-based surveys, whether administered by interviewer or self-administered by a

respondent in a postal survey. Section 4.4 discusses the steps for developing survey specifications that provide more detailed information to facilitate questionnaire programming into an electronic data capture system for surveys using a form of electronic data capture. It will focus on interviewer-administered surveys, but will also provide guidance on development of specifications for web surveys. Section 4.5 provides step-by-step guidance for translation of the questionnaire into other languages, while Section 4.6 provides an overview of several questionnaire evaluation and testing approaches, informed by the QDET framework.

17. Sections 4.7 – 4.8 provide an overview of systems implementation for survey instruments. Section 4.7 reviews the critical elements of an electronic data capture system. Section 4.8 provides guidance on programming and testing of systems.

18. Lastly, Section 4.9 identifies emerging approaches in questionnaire design, translation, and systems implementation as of the time of this publication.

4.2. Questionnaire content

19. While there are numerous publications on the development of questions (Bradburn et al. 2004; Converse & Presser, 1986; Fowler, 1995; Groves et al., 2009; Willimack et al., 2004), this section provides an overview of the most critical elements in questionnaire development.

4.2.1. Research objectives and analysis plan

20. A well-thought-out set of research objectives and associated analysis plan is essential for creating an effective and efficient questionnaire. Research objectives are a critical component of scope determination (see Chapter 2) and form the framework for the analysis plan (see Chapter 9). While the data analysis stage is one of the later stages in the survey lifecycle, there are several reasons why developing an analysis plan is a critical step early in the questionnaire content development process. Specifying the analysis plan more precisely identifies the variables for which data must be collected, along with the most appropriate format (e.g., categorical, continuous, etc.). Alignment of specific questionnaire items with corresponding analysis variables also provides an opportunity to retain only those questions for which there is a proposed analysis purpose, reducing both respondent and interviewer burden as well as survey cost. This strategy also ensures that critical items are not omitted from the final questionnaire and that all included items are directly linked to the main research objectives and allows researchers to identify the most critical topics that deserve more attention in the survey instrument, thereby enhancing the relevance and focus of the survey.

21. [[Case study requested on alignment between research objectives and survey questions.]]

4.2.2. Determining questions

22. Following the definition of the research objectives and the analysis plan, the next step in questionnaire design is to select survey questions that measure the relevant concepts. When determining the wording of specific survey items for the questionnaire, researchers may elect one or a combination of several strategies: (1) reuse questions without any adaptation, (2) adapt existing questions, and (3) write new questions (van de Vijver et al., 2003).

4.2.2.1. Reusing questions

23. Reuse of existing survey questions has several advantages. When collecting trend data, it is crucial to maintain consistency in the wording and format of questions across different iterations of the survey, so any changes in responses over time reflect true changes in the population, rather than changes in the questionnaire (Krosnick & Presser, 2010). Produce accurate trend data is particularly important for national

statistical organizations, when collecting comparable data for benchmarking or studying trends across surveys is the primary objective. Benchmarking is especially useful when conducting a sub-national survey or a survey with a smaller sample size, with the goal to assess how the target population compares to the national population.

Box 4.3. What does it mean for survey items to be "validated"?

- Face validity assessed through expert review of the survey items to ensure they appear to measure what they are supposed to measure
- Content validity assessed by consulting with subject matter experts and reviewing the literature to ensure that the survey items cover all relevant aspects of the construct being measured
- Cognitive interviews conducted with a small sample of respondents to understand how they interpret the survey questions (See Section 4.1…
- Pilot testing conducted with a small, representative sample of the target population
- Reliability testing conducted to assess the consistency of the survey items by calculating reliability coefficients, such as Cronbach's alpha
- Factor analysis conducted to examine the underlying structure of the survey items.
- Criterion validity assessed by comparing the survey results with other established measures or outcomes to ensure they align

24. Use of validated survey items when available that have been tested and proven to be reliable and valid in measuring the construct of interest can enhance the accuracy and credibility of the data collected (Fowler, 1995; Gehlbach & Brinkworth, 2011). For example, Northwestern University's Institute for Policy Research, in collaboration with a consortium of international partners, developed and validated a set of water insecurity experiences (HWISE) scales through a research study in 23 LMICs.[3] Note, however, that scales which have been validated in one context may not be valid and reliable in another context and may benefit from additional validation in a contextually-appropriate environment before fielding. For example, a review of different sets of survey questions used to collect food consumption data revealed methodological concerns that could negatively impact data quality in several of the approaches (Smith et al., 2014).

25. There are several online resources available for existing questions, including data archives such as the University of Michigan Inter-university Consortium for Political and Social Research (ICPSR), USAID's Data Documentation Library (DDL), the International Household Survey Network (IHSN), the repository of measurement and survey design resources hosted by J-PAL, World Bank Open Data, National Center for Health Statistics Q-Bank, and ZIS Open Access Repository for Measurement Instruments hosted by GESIS and Fielded survey questionnaires are publicly available on most national statistical organization websites and are another important source of survey questions. Questionnaires from long-running large-scale international surveys, such as the Demographic and Health (DHS) Surveys, UNICEF's Multiple Indicator Cluster Survey (MICS), the World Bank's Living Standards and Measurement Surveys (LSMS), and the European Social Survey (ESS) are also often publicly available online. If a question is to be reused,

[3] See Water Insecurity Experiences (WISE) Scales - Northwestern University for more details.

confirm that no contextual adaptations will be required, particularly if the question has never been implemented in a survey in the anticipated target population.

4.2.2.2. Adapting questions

26. Adaptation of existing questions involves deliberate alteration of the content, format, response options, or visual presentation of any part of the question to make the resulting survey item more suitable for another sociocultural context or a particular population. This is often necessary when reusing questions from other contexts. For example, the World Bank's Living Standards and Measurements (LSMS) survey collect data on household consumption and expenditure across a wide variety of contexts. However, common household items, types of food consumed, and other such items of interest vary and corresponding survey items require significant adaptation from country to country. Similarly, units of measurement often require adaptation when existing survey questions are used. Adaptation of questions should occur in close coordination with knowledgeable members of the target population to ensure contextually-appropriate terminology in questions.

4.2.2.3. Developing new questions

27. Lastly, a survey instrument may require drafting of new questions to achieve the research objectives or to be suitable for the target population. If new questions are to be drafted, first define the broad topic and then break it down into more manageable subtopics. Subtopics are further subdivided until they reach the level of single issues or concepts that can be articulated as survey questions and meet the needs identified in the analysis plan.

28. Focus groups are an effective tool for drafting initial questionnaire items. By bringing together a diverse group of participants in one or more focus groups, researchers can explore a wide range of perspectives and experiences related to the survey topic (Presser et al., 2004). During these discussions, participants can provide feedback on the clarity, relevance, and wording of potential survey questions. This collaborative process helps identify any ambiguities or biases in the questions and ensures that the language used resonates with the target population (Morgan, 1996). In-depth, semi- or unstructured interviews with key informants can play a similar role in the questionnaire design process, providing researchers an opportunity to explore concepts relevant to the topic(s) of interest (Axinn & Pearce, 2006; Gerber, 1999). These qualitative methods can highlight important themes and topics that might otherwise be overlooked, guiding the prioritization of content in the questionnaire early in the design process.

29. There are two primary types of survey questions: those that gather factual data, and those that gather subjective data. Questions that gather factual data include those on objectively verifiable facts and events, such as household demographics, housing characteristics, and household member experiences and behaviours. While the range of topics is wide, the commonality is that there is an accurate response to these types of questions that, theoretically, can be verified. In contrast, questions that gather subjective data measure respondents' knowledge and perceptions, feelings, attitudes, and judgments. While consistency in responses can be assessed, there is no external method to assess the accuracy of such questions.

30. Although factual and subjective survey questions differ in verifiability, the underlying definition of a "good question" is similar for both:[4]

[4] See https://www.abs.gov.au/statistics/standards/abs-forms-design-standards/2023 for a comprehensive manual on best practices for questionnaire design. See also Fowler (1995) and Sudman & Bradburn (1982) for further guidance on writing good survey questions.

- All respondents have a shared understanding of the meaning of the question
- Respondents are asked questions to which they should know the answers
- Respondents are able to answer questions using the response categories provided by the question
- Respondents should be asked questions they are willing to answer accurately

4.2.3. Tips for writing questions

4.2.3.1. General tips

31. Following are a number of general tips when drafting questions, regardless of the question type, along with some examples of poor-quality survey questions and improved versions:

- Use simple, familiar words (avoid technical terms, jargon, and slang)
- Use simple syntax and language
    - Writing questions at a fourth-grade reading level is often recommended
- Avoid words with ambiguous meanings and aim for wording that all respondents will interpret in same way
- Aim for wording that is specific and concrete (as opposed to general and abstract)
- Avoid leading questions that push respondents toward an answer
    - Original: *Do you agree that the new water filtration system has greatly improved the quality of life in our village?*
    - Revised: *How would you rate the impact of the new water filtration system on the quality of life in our village? Would you say it has had a very positive, somewhat positive, somewhat negative, or very negative impact?*
- Ask about one thing at a time; avoid double-barreled questions
    - Original: *Do you have access to clean drinking water and reliable electricity in your home?*
    - Revised: *Do you have access to clean drinking water in your home? Do you have access to reliable electricity in your home?*
- Avoid *questions* with negatives
    - Original: *Do you agree that there is no clean water access in your community?*
    - Revised: *Is there clean water access in your community?*

4.2.3.2. Tips for factual questions about non-sensitive behaviours

32. When asking factual question about non-sensitive behaviours, consider the following additional guidelines:

- If the question is closed, include all reasonable options. If the actual answer not listed, respondents likely to underreport it.
- Use recall cues, like lists of specific activities, when forgetting is a concern
    - Original: *How many organizations do you belong to?*
    - Revised: *How many organizations do you belong to – for example, unions, churches, community organizations?*

- Make the question as specific as possible; make it clear who and what is covered by the question and what time period
    - Original: *When do you usually leave home to go to work?*
    - Revised: *During the last week, when did you usually leave home to go to work?*
- When asking about the number of events in a specific time period, pick a duration that is appropriate for topic
    - Hospital stays – six months
    - Doctor visits – a couple of weeks
    - Dietary recall – one or two days
- Use aids like event or life history calendars that provide a visual timeline of significant events, aiding respondents in accurately remembering and reporting past behaviours and experiences (Axinn et al., 1999; Freedman et al., 1988)
- Use words that virtually everyone understands and define special terms
    - Original: *Do you procrastinate?*
    - Revised: *Do you put off until tomorrow things you could have done today – that is, procrastinate?*
- Increase the length of questions by adding memory cues
    - Original: *Did you ever use any healthcare service even once in the past six months?*
    - Revised: *There are a lot of different types of healthcare services. By healthcare services, we mean visits to doctors, nurses, clinics, and similar providers, as well as telehealth consultations and wellness check-ups. Did you ever use any healthcare service even once in the past six months?*
- Consider unfolding brackets for more difficult estimates
    - Original: *How many meters do you generally walk for water?*
    - Revised: *How many meters do you generally walk for water? Would you say more than 500 meters or less than 500 meters? (If less than 500 meters: More than 250 meters or less than 250 meters?) (If more than 500 meters: More than 1000 meters or less than 1000 meters?)*

4.2.3.3. Tips for subjective questions

33. When designing questions to collect subjective data, consider the following guidelines:

- Use balanced items rather than unbalanced items, where one of the responses may be ambiguous.
    - Original: *Are you in favor of the changes to the healthcare system recently enacted by the government?*
    - Revised: *Would you say you strongly favor, favor, oppose, or strongly oppose the changes to the healthcare system recently enacted by the government?*
- Use bipolar scales with alternative positions as end points rather than unipolar scale, which has varying levels of the same construct, without a conceptual midpoint and with a zero at one end
    - Original: *To what extent do you feel that the healthcare services in your community meet your needs? Would you say not at all, to a small extent, to a moderate extent, or to a great extent?*

- Revised: *How satisfied are you with the quality of healthcare services in your community? Would you say very dissatisfied, somewhat dissatisfied, somewhat satisfied, or very satisfied?*

- Put general item before specific items on same issue to accommodate contextual effects (Schwarz & Strack, 1991)
  - First question*: Taken all together, how would you say things are these days? Would you say that you are very happy, pretty happy, or not too happy?*
  - Second question: *Taking things all together, how would you describe your marriage? Would you say that your marriage is very happy, pretty happy, or not too happy?*

- When asking questions with differing levels of popularity, ask the least popular question first

- Use between five- and seven-point scales, and verbally label every point (Alwin & Krosnick, 1991)

- Avoid check-all-that-apply items; ask for an answer for each item (e.g., yes/no)

- Avoid use of the Likert scale (agree/disagree format)
  - Original: *Do you strongly agree, agree, disagree, or strongly disagree that the current food prices are making it difficult for you to buy enough food each week?*
  - Revised: *To what extent are the current food prices making it difficult for you to buy enough food each week? Would you say a great deal, somewhat, not very much, or not at all?*

4.2.3.4. Tips for sensitive items

34. When designing questions to collect sensitive information, tailor the questionnaire mode used to collect sensitive data to the interview context where possible. The optimal mode to collect high-quality data on sensitive topics differs significantly by topic and context (de Jong, 2016). A mixed mode design may be optimal, where the majority of a questionnaire is administered by the interviewer, but sensitive items are in a self-administered module to be completed privately by the respondent.[5] Regardless of mode, however, consider the following guidelines when designing questions that collect sensitive data:

- Use open-ended questions rather than closed-ended questions for obtaining frequencies of sensitive behaviours to lessen perceptions of judgment

- Avoid using a middle category; respondents may erroneously infer that the middle category is typical (mean) response and extreme categories are unusual

- Use showcards so respondents do not have to verbalize sensitive terms

- Deliberately begin the question with a statement indicating the behaviour in question is common, to put the respondent at ease:
  - Example: *Even the calmest parents get angry at their children sometimes. Did your child(ren) do anything that made you yourself angry during the last seven days?*

- Presuppose behaviour to put the respondent at ease:

[5] See Chauchard, 2013; Hewett et al., 2004; and Mensch et al., 2003 for approaches to implementing self-administered modules in low-literacy settings. If an alternate mode is not possible, consider the Random-Response technique (Lara et al., 2004) or Item-Count technique (Miller, 1984).

- Example: *When was the last time you smoked a cigarette?*

- Use authority to justify the topic of the survey question
  - Example: *Many doctors now think that drinking can reduce heart attacks. Have you ever had a drink of alcohol, even once in your lifetime?*

- Use forgiving wording
  - Example: *Did things come up which prevented you from voting or did you manage to vote this time?*

- Start with lifetime questions
  - Example: *Did you ever even once smoke a cigarette?*

- Embed sensitive items among other items to reduce the item's prominence
  - Example: *Include a question about abortion among other gynecological procedures*

- Use repeated measurement (e.g., diaries or panel surveys) to study sensitive topics; respondents will get used to discussing these issues

4.2.4. Response options

35. Survey questions can be open-ended or closed-ended. In **open-ended questions**, respondents are not offered a response option; rather, they are invited to answer in their own words. In an interviewer-administered survey, the response may be captured in one of two ways. In the verbatim open-ended approach, the interviewer enters into the survey instrument the respondent's verbatim words. The resultant data from this approach can be challenging because they require significant time and effort to analyze, as responses are often lengthy and varied. Additionally, the qualitative nature of the data can make it difficult to synthesize and categorize responses into actionable insights. In the field-coded open-ended approach, the interviewer has a list of response options in the survey instrument and selects the appropriate closed-ended response, based on what the respondent says. The final category on the list of options is often "other, specify", which will likely result in at least a few verbatim responses that will require post-data collection processing.

36. In **closed-ended questions**, respondents are either offered a list of response options following the survey question, or otherwise receive indication that the response should be a number (e.g., number of trips to the doctor in the past month). Response options can vary from a simple "yes/no" to an extensive list of options, such as a list of government programs accessed in the past year. Show cards, which are cards that have the response options for a survey question printed on them can be a useful tool in face-to-face interviewer-administered surveys when there are more than a few response options. Using show cards in a survey interview can help respondents easily visualize and choose from a large number of response options, reducing confusion and improving the accuracy of their answers. However, the utility of showcards will vary depending on the literacy of the target population.

37. While a guiding principle is to include survey questions that respondents are willing and able to answer, there will be some questions for which respondents may struggle to respond because they cannot determine the correct response or because they do not want to reveal the correct response. As discussed in Chapter X, many surveys require approval from an institutional review board (IRB) or other ethical review, which in turn often requires interviewers to obtain informed consent from respondents before beginning the interview. An important component of informed consent is a statement that respondents can skip any question that they would like. Thus, it is important in both open- and closed-ended questions to allow for

respondents to indicate that they would like to skip a survey question, either by stating “don’t know” or by refusing. Therefore, both “don’t know” and “refused” are often included as response options to each survey question, although they are not customarily read aloud to the respondent in an interviewer-administered questionnaire.

4.2.5. Interviewer guidance

38. Interviewer instructions within the questionnaire ensure that interviewers understand the flow of the questionnaire and how to administer each question. Instructions may indicate to interviewers to “read all options aloud”, “select all response options that apply”, or “enter respondent’s birth year as a four-digit number”. Such instructions may be visually differentiated from the survey questionnaire using a specific format, as discussed further in Section 4.4. Additionally, as discussed in Chapter 6, the questionnaire may also include guidance for interviewers to apply conversational interviewing principles when clarification is needed, wherein interviewers have flexibility in providing a definition or other clarity to the respondent while minimizing the introduction of systematic bias in the interview (Conrad & Schober, 2021).

4.2.6. Questionnaire scope

39. One of the key challenges at this stage of questionnaire design is managing the scope of the survey through decisions with stakeholders. Decisions, difficult at times, will need to be made about what to retain and what to drop, based on the relevance to the research objectives and the analysis plan. As noted previously, the questionnaire design process is iterative and may require revisions to the initial objectives and anticipated analysis, depending on the ultimate length of the draft questionnaire and consideration of what is feasible in terms of the survey budget and burden to the respondent. Lengthy questionnaires can also have a significant impact on data quality. For example, evidence from a study in Liberia and Malawi suggests that in a survey with an average time of 2.5 hours, an additional hour duration resulted in a significant item non-response rate as well as a decline in response accuracy (Jeong et al., 2023).

4.2.7. Question organization

40. Household questionnaires are often developed in a modular format, with questions organized thematically (e.g., by housing characteristics, children’s nutrition, etc.) and by level of respondent (e.g., household level, primary female adult decision-maker, etc.). This is effective both from a design perspective to facilitate question organization as well as from the respondent perspective, to reduce cognitive burden that would result from significant topic switching during the course of the survey interview. It is best practice to begin the survey interview with a set of questions that are easy for the respondent to answer and are not sensitive. This helps to motivate the respondent and provides an opportunity to build the rapport between the interviewer and the respondent.

41. The order in which survey questions are presented both within and between thematic sections can also influence respondents' answers, with context effects occurring when responses to a prior question on a questionnaire affect responses to a subsequent one. For example, a common approach to asking about household expenditure is to first ask respondents whether they own each of a number of household items, and then ask about the value of each. While a more streamlined approach would be to ask about the value of each item to which the respondent affirmed ownership, the respondent would quickly learn that responding ‘yes” leads to an immediate follow-up question and may be more likely to underreport ownership so as to avoid the immediate follow-up question. However, by asking about ownership to all questions first, and then asking about value, there is a lower likelihood of under-reporting of ownership and better data quality. This type of screening question is recommended particularly when the questionnaire is lengthy and there is an opportunity for the respondent (and the interviewer) to learn shortcuts that would shorten the interview length.

42. Interestingly, there is evidence that the impact of question order can vary with age, as older respondents with lower working memory are less influenced by preceding questions (Knauper et al., 2007), and there is growing evidence of variance in both question and response category order effects by other demographic characteristics as well (Rasinski et al., 2012). Lastly, the placement of sensitive questions should also be considered, with guidance suggesting placement at the end of the questionnaire to provide ample opportunity to build the relationship between the interviewer and the respondent in interviewer-administered questionnaires and to increase the overall completeness of the survey, in the event that the respondent requests to end the interview upon administration of the sensitive items.[6] There is always the possibility for a context effect to occur in a questionnaire. When determining the order of the questions, consider whether there are reasons for concern based on the question content and target population and include explicit assessment of any potential context effects in the evaluation and testing of the questionnaire.[7]

4.2.8. Advance translation

43. An initial step to enhance the translatability and cross-cultural portability of the source questionnaire, referred to as "advance translation", is recommended before finalizing the source questionnaire and beginning the translation process as outline in Section 4.5 below (Dorer 2020, 2023). "Equivalent translations in a cross-national study require as a first step a source questionnaire that is suitable and possibly annotated for translation and cross-cultural transfer – which calls for cross-cultural cooperation during the development of the questionnaire (Smith, 2003)" (Behr 2017: 582). When translation of a questionnaire into one or more target languages will be required, this perspective ideally becomes an integral part of the source questionnaire development. If the source questionnaire is developed with subsequent translation in mind, the translation process and thus the translation results can be improved, later translation queries or issues minimized or avoided, and thus the overall quality of survey data improved. Advance translations were recommended by Harkness and Schoua-Glusberg as early as 1998 (Harkness & Schoua-Glusberg, 1998) and have been applied by the European Social Survey (ESS) since 2009 (Fitzgerald & Zavala-Rojas, 2020), by several surveys administered by Eurofound since 2013, and the New York City Housing and Vacancy Survey (NYCHVS) among others (Waickmann & Corbett, 2022). The efficacy of advance translation was confirmed in an empirical thinking-aloud study by Dorer (2020, 2023).

44. Advance translation consists in translating a pre-final version of the source questionnaire and identifying translation problems that could be avoided by modifying the source text before the source questionnaire is finalized and the more extensive translation process starts.[8] Advance translation consists of translating a pre-final version of the source questionnaire by a translation team, carrying out the steps of Translation, Review and Adjudication, from the TRAPD approach as explained below in Section 4.5.2, in which two independent parallel Translations are carried out and discussed in a Review meeting, in which both trained and experienced questionnaire translators and survey experts and/or social scientists determine the best translation. Any issues that could not be resolved in the Review step are subsequently addressed in the Adjudication step. When this team approach is carried out for Advance Translation, the primary

[6] See also https://www.abs.gov.au/statistics/standards/abs-forms-design-standards/2023/general-forms-design-principles-question-structure#carefully-consider-where-sensitive-questions-are-placed for further guidance on placement of sensitive questions

[7] See also Schwarz et al. (1992) and Schwarz (1999) for further discussion on questionnaire context effects.

[8] An alternative approach to Advance Translation is a Translatability Assessment (see Acquadro et al, 2018).

objective is to report on any translation problems that are linked to the source questionnaire, identifying mainly: (a) items lacking clarity or information of the source text, (b) elements of the source questionnaire which will be difficult or impossible to translate into the target languages and/or cultures due to either linguistic or cultural issues, and (c) any other design problems in the source questionnaire detected in the Review discussion. Solutions to tackle the source questionnaire problems detected are mainly (a) rewording the source questionnaire, and (b) adding translation notes to clarify the source text and providing translation briefs.

46. Interdisciplinary translation teams consisting of trained and experienced questionnaire translators and survey experts / social scientists are best poised to carry out advance translations (see Section 4.5.2). Ideally all target languages and cultures of the cross-cultural survey should be covered to ensure all potential translation problems can be detected and resolved before the source questionnaire is finalized. If this is not possible, e.g., because of multiple target languages or financial constraints, the recommended approach is to select representatives languages to cover all linguistic and cultural groups as much as possible, e.g., one Romance language from Europe, one Bantu language from Africa and one Asian language.

47. Ideally, the development of the source questionnaire already includes cross-cultural input (Fitzgerald et al, 2011, Fitzgerald & Zavala-Rojas, 2020). For example, it should address elements that may be problematic or impossible to ask in certain countries, such as questions about sexual behaviour, or other cross-cultural differences, like "walking several blocks" in countries where the concept of a "block" is unknown, as noted in Section 4.2.8. If such considerations are incorporated into the source questionnaire, the additional steps to enhance the source text's translatability may be less necessary. Where the source questionnaire was developed without any cross-cultural input, steps like advance translation become more important in preparation of the final translation process.

48. Given the importance that advance translation can plan, some level of assessment of the source questionnaire for its fitness to be translated and fielded in all planned target languages and cultures is recommended. The translation industry recognizes the role that the source text plays in the translation process and the quality of the resulting translations, and source text issues are frequently detected in the translation process (Hauck 2004). If these issues are discovered only after the source questionnaire is finalized and cannot be modified, it may be too late to avoid certain translation problems and errors, if it is, e.g., detected that certain terms or concepts cannot be fielded in certain cultures or cannot be expressed in certain languages. Simply screening a source text often misses these issues.

Box 4.4. Example of the efficacy of advance translation from the ESS

- Item pre- advance translation: "To what extent do you receive help and support from other people when you need it?"
- In the advance translation, the translation teams asked whether the help or support was financial or personal in nature, so in the final source questionnaire, a translation note was added explaining that "help and support" was to be understood as "emotional or material."
- Without this clarify, some of the 30+ language versions may have translated help and support in the financial sense, and others in a personal sense, and the resulting survey data would not have been comparable (see Dorer, 2020 and Dorer, 2023).

4.2.9. Cross-cultural considerations

49. When designing questions for administration across different national, cultural, or lingual groups with the intention of comparative analysis, cultural and linguistic differences can impact respondents' cognitive processes, particularly when responding to subjective questions (Peytcheva, 2020; Schwarz et al., 2010; Tourangeau et al., 2000; Uskul et al., 2010; Uskul & Oyserman, 2006). Cultural differences also influence what people pay attention to and how they organize information. For example, Ji and colleagues (Ji et al., 2000) demonstrate how Chinese vs. American students differ in their recall strategies for survey items when asked about observable vs. unobservable behaviours. In terms of reporting information, social desirability bias has been also found to be positively associated with societies where collectivism, rather than individualism, is the dominant cultural norm (Bernardi, 2006; Bond & Smith, 1996; Lalwani et al., 2006; Triandis, 1995). Other studies have examined the impact of cultural frames on acquiescence (the tendency to choose "agree" or "yes" responses), extreme, and middle category response styles on survey responses (see Yang et al. (2010) for a detailed review and discussion). If subjective questions are included, particularly in a survey collecting data for cross-cultural comparative analyses, consider methodological analyses post-data collection to determine whether there is evidence of measurement error among those types of questions that the literature suggests are most vulnerable.

4.3. Technical design for paper-based questionnaires

50. As noted earlier in the chapter, the majority of surveys are administered in modes using electronic data capture. However, there are instances where paper-and-pencil questionnaires are used by interviewers in face-to-face surveys (see Section [[6.X]]). In other surveys, paper questionnaires are sent through postal mail for respondents to self-complete and send back to researchers. This section covers both interviewer- and self-administered paper questionnaires. The technical design of a paper-based questionnaire, which includes the format, layout, and visual presentation of survey questions, can impact usability, and poor technical design can lead to measurement errors and increase both interviewer and respondent burden (Hansen et al., 2016).

51. [[Case study requested on use of a paper-based questionnaire in a face-to-face survey.]]

52. A paper-based self-administered questionnaire must be straightforward for the respondent to self-administer, with respondents easily able to recognize instructions (such as "Select one") and to read questions, navigate correctly through the instrument, and enter responses (Dillman et al., 2005; 2009). Instructions should appear where they are needed, such as "Start here" before the first question and response entry instructions (e.g., "Tick all that apply") after the question text. Response categories should be displayed in the order of their likely occurrence. In addition, instructions to skip questions should be avoided or used sparingly in paper self-administered instruments because they can lead to response errors. Complexity in interviewer-administered paper questionnaires should be as limited as possible as well. It is also important to evaluate the readability of colour, graphics, images, maps, and icons in paper questionnaires.

4.4. Questionnaire specifications preparation for electronic data capture

53. For surveys using electronic data capture, either in-person (CAPI) or by phone (CATI), the next step after finalizing questionnaire content is to prepare questionnaire specifications. The questionnaire specifications are documentation that provides the survey programmer with clear guidelines to develop a draft of the programmed survey instrument with minimal need for clarifications on the questionnaire's content and its programming. Skipping this step and moving directly from an unstructured questionnaire to programming can lead to errors, inefficiencies, as well as an unnecessary and increased back and forth

between the survey team and the survey programmers. To ensure accuracy and efficiency of the programmed instrument, the survey team should invest time in creating detailed questionnaire specifications.

54. [[Case study requested on effective development of questionnaire specifications as an intermediary step between questionnaire design and CAPI instrument development.]]

55. Begin by developing a template for the survey specifications. The template should include conventions for conveying survey content, logic and instructions clearly. The template and conventions used may vary from one survey to another; therefore, it is essential to document these at the start of the specifications for clarity[9]. Examples of common conventions include: (i) using different colors for programming instructions, respondent-facing text, and interviewer instructions, (ii) capitalizing programming instructions, and (iii) inserting variable names at the beginning of each question (Couper et al., 2000; Hansen et al., 2016).

56. To ensure that the specifications are comprehensive, consider including the following key elements in the specifications:

- **Question numbers:** Sequential numbering at the start of each question is useful in interviewer-administered surveys, as it can improve interviewer training and help interviewers communicate progress and issues with supervisors more effectively throughout data collection. However, for self-administered surveys, sequential numbering is not necessary and can be omitted without impacting the survey process.

- **Variable Names:** Each item in a survey is assigned a unique identifier known as a variable name, which is essential for storing and referencing the data collected from respondents. Variable names may be either the question number or be descriptive of the question, as they play a crucial role in quality control and data analysis by providing a concise reference to specific questions. For example, a question asking for a respondent's age might have the variable name "age". When assigning variable names, avoid variable names that begin with a number and avoid the use of special characters like spaces, periods, or symbols (e.g., “#”, “!”, “?”, or “@”) that may cause issues in data processing software, with the exception of underscores (“_”).

- **Question Text:** This field will display the exact wording and formatting of the question that interviewers will see. An example of question text and formatting might be, " What is the **main** source of drinking water for members of your household?". In this case, the programmer would insert the question text verbatim and apply formatting to underline and bold the word “main”.

- **Response Option Codes:** Response options should be coded consistently throughout the questionnaire. For instance, all yes/no questions could be coded with the same values (e.g., “1” = “Yes”; “2” = “No”). Similarly, consider standardizing codes used for other frequently used response options such as “Other (specify)”, “Don’t know”, and “Refused” (e.g., “-88” = “Other (specify); “-98” = “Don’t know; “-99” = “Refused”). To avoid confusion, the numeric codes do not need to be displayed on the phone, tablet, or computer screen during the interview, however, once selected, the responses will be assigned the appropriate code. The survey team should

[9] Example: National Center for Health Statistics: Rapid Surveys System. RSS-1 Questionnaire Programming Specifications. Hyattsville, Maryland. 2023

consider using negative numbers for response options to numeric questions to prevent valid responses, such as an age of "98" from being misinterpreted as "Don't know".

- **Interviewer Instructions:** For ease of programming, interviewer instructions should be visually differentiated from question text and should be displayed according to the agreed-upon conventions. For instance, an interviewer may be instructed to never read response options "Don't know" or "Refused". The specifications would include interviewer instructions indicating that these response options should not be read (i.e., -98. Don't know (Do not read) and -99. Refused (Do not read)). In some instances, the survey team can use capital letters to indicate text that should not be read aloud. However, this approach may not be suitable in all countries, as some languages do not use capital letters, which can result in translation inconsistencies or mismatches. In other contexts, square brackets may be used to indicate text which should not be read aloud.

- **Enabling Conditions / Skip logic:** Enabling conditions or skip logic refers to the logic used to determine if certain questions are displayed to a respondent based on their previous answers. In some contexts, this is also referred to as "filtering" or "routing". When specifying enabling conditions or skip logic in the specifications, include both the variable names and response option codes as part of the syntax. To reduce the risk of programming errors, consult with the survey programmer to determine if the data collection software utilizes enabling conditions, skip logic, or both.

- **Validation Conditions:** These are the criteria that a response should meet before being accepted by the survey instrument. For example, a multiple-choice question may be designed to accept multiple responses (i.e., all that apply), with the exception of "Don't know" or "Refused", which may only be selected as a standalone response. e. Similarly, a question like "age", may require a range check to ensure that the response is valid. For example, a range check may be included in the questionnaire specifications to verify that a respondent's age is less than or equal to 99 in target populations where values of 100 or more are highly unlikely.

- **Pre-filled data:** Consider using a standard method to indicate question wording that will be filled using data provided earlier by the respondent, such as a household member's name. Pre-filled text can also be used to facilitate recall. For example, in a question asking about an experience in the last 12 months, the specification may include "In the past 12 months, since [MONTH P1YEAR], have you…", with the text in brackets populated to display the current month and the previous year, reminding the respondent of the specific time period in question.

- **Other programming logic:** Consider including additional programming notes relevant to the survey programmer. This may include comments on (i) how many questions should be displayed per screen, (ii) whether to randomize the order of questions or response options for respondents, or (iii) the need to display survey aids for specific questions.

4.5. Translation

57. If the questionnaire is first produced in one language (i.e., the source language), but will be administered in another (i.e., the target language), a written translation of each target language will generally be required.[10] This section provides guidance on the steps recommended to produce a high-quality translation of the questionnaire in each target language used in the survey.

[10] For exceptions to a written translation of a target language, see further discussion in Section [[4.5.X]].

4.5.1. The importance of translation

58. Translation is the conversion of text from one language into another in written form. When respondents with different mother tongues are to be surveyed, it is necessary to translate the questionnaire before it is administered in ensure comparability both between different languages as well as within the same language. Without written translations, interviewers have to translate the questionnaire themselves during the interview, often resulting in "on-the-fly" translations. While sometimes necessary, as discussed in more depth below, this approach can lead to slight variations in phrasing each time the interviewer asks a question and thus compromise the standardization of the interviews (Blom et al., 2006).

59. For survey translation, the language in which the questionnaire is initially developed is the "source language", while each language to which the questionnaire is translated is the "target language". Once the source questionnaire has been finalized, ideally having passed an advance translation to enhance its translatability, it is ready to be translated, although, depending on the timeline and other project management constraints, issues identified in the translation process of the final source questionnaire may result in more significant revisions to the source questionnaire itself.

60. Not only do the questionnaires have to be translated before being administered, but these translations also need to be of high quality. Producing high-quality translations is essential to ensure that data from each translated version of a questionnaire are comparable. All translations must guarantee that "the same questions" are asked across all language versions. This approach is also known as "Ask-the-same-question" (ASQ) (Harkness et al., 2010a).

61. For some surveys, the ASQ approach plays an integral role from initial translation of the source question all the way through to the stage of data dissemination. For example, the European Social Survey (ESS) will remove variables from its integrated dataset if it is known that the data is based on erroneous translation, as noted in **Box 4.5**, where the question asks about the acceptability of an immigrant of a **different** race or ethnic group becoming the respondent's boss or marrying a close relative.[11] In one ESS country, the question was translated this to ask about "the **same** race or ethnic group". As this would mean that the respondents in this country would not answer the same question as in the English and all

Box 4.5. Example of a translation error

ESS7:

Now thinking of people who have come to live in [country] from another country who are of a **different** race or ethnic group from most [country] people. Using this card, please tell me how much you would mind or not mind if someone like this ...

D10 ...was appointed as your boss?

D11 ...married a close relative of yours?

AUSTRIA/GERMAN:

Wenn Sie nun an Zuwanderer denken, die aus einem anderen Land nach Österreich gekommen sind, und **derselben** Volksgruppe oder ethnischen Gruppe angehören wie der Großteil der österreichischen Bevölkerung, wie sehr würde es Sie stören, wenn so jemand ...

D10 zu lhrem Vorgesetzten gemacht würde?

D11 eine/n enge/n Verwandte/n von Ihnen heiraten würde?

[11] See the text box for an example of a translation error resulting in extracting a variable from an integrated dataset (European Social Survey, Round 7: ESS7 - integrated file, edition 2.3 | ESS - Sikt / DOI: 10.21338/ess7e02_3)

other target language versions, the country's data could not be used for cross-cultural comparisons and the data from the country for this variable was removed from the integrated dataset.

4.5.2. Determining target languages

62. The definition of the target population will largely determine the target language(s) for the survey. In monolingual countries, such as Korea, where the target population is linguistically homogeneous, no translation is necessary if the source questionnaire is produced directly in Korean. Translation into only one target language (i.e., Korean) will likely be necessary if the source questionnaire is produced in a different language (e.g., English).

63. However, many countries are multi-lingual. Some countries have several official national languages, which are often all identified as target languages. Targeting certain minority groups may be important and relevant to include in the final survey data, even if those minority languages are not official national languages. With increasing migration, globalization and demographic changes, systematically excluding certain groups from a survey may lead to non-response bias because of language barriers (Beullens et al 2017; Lipps & Ochsner 2018). Some surveys use a minimum threshold of speakers of a language. For example, the European Social Survey (ESS) requires that fieldwork is conducted (and thus the source questionnaire translated) in all languages spoken by 5% or more of the resident population as their first language (European Social Survey 2024). Among United States Agency for Development (USAID) Feed the Future countries, however, the threshold for translation is 10% of the target population (ICF, 2024). The decision of which languages to include in the fieldwork may also be a political decision, particularly in areas where there is civil unrest or regional conflict. Another reason to use certain languages may be trend data: if trend data have been gathered by fielding a survey in certain languages, languages used to collect these data should be maintained over time.

64. Some languages are oral only, or are written only rarely and require transliteration (the process of writing a word using the closest corresponding letters of a different alphabet or script). When a target language is used only orally, both translation of the source questionnaire as well as administration of the target questionnaire can be challenging, as translators may struggle with the transliteration process, and interviewers may struggle to read the subsequent translation. In such instances, additional time for both the review of the transliterated questionnaire as well as for interviewer training will be necessary. If a written translation for a target language cannot be produced, and "on-the-fly" oral interpretation will be required in the field, consider providing a list of key terms from the questionnaire to interpreters to ensure at least a minimum level of consistency in translation across languages during data collection, and include a section on on-the-fly translation during interviewer training (see Chapter [[6.X]]).

65. In the case of multi-country surveys, where languages are spoken in more than one country (often referred to as "shared languages"), such as French, English or Chinese, we recommend that these language variants are taken into consideration as much as possible. Respondents should be addressed in a language that is as close as possible to their everyday language. Thus, we recommend a unique version of the target language questionnaire for each country (e.g., a French version for France, and separate version for each of Canada, Belgium, and Cameroon). There are different approaches to developing these language versions: (a) following the full translation process, as outlined below, which is recommended where the language versions differ substantially, or (b) adapting or localizing other versions of this language (e.g., using the French version developed for France and localizing it for use in Cameroon as well as each additional French-speaking country). Such an approach may be acceptable where the language versions are very similar or may differ only in factual terms, e.g., administrative or political terms, such as education levels, political parties or the names of certain government institutions or ministries and would, for instance, be

recommended for very recent migrant groups from one particular country which have not yet formed a new language variant in their new host country.

4.5.3. The TRAPD approach to translation

66. Determine the translation method or process before starting the translation process. The recommended approach for multilingual survey projects is the team or committee approach, ideally in its elaborate form called "TRAPD" (Harkness 2003, Harkness et al 2004, Harkness et al 2010b, Behr & Shishido 2016). The TRAPD approach involves: (a) an iterative translation process and (b) preparation by an interdisciplinary team, including experienced and trained professional translators and social scientists/survey experts. **Figure 4.*3*** illustrates the TRAPD approach to translation, defined as follows:

T: Two or more independent parallel translations

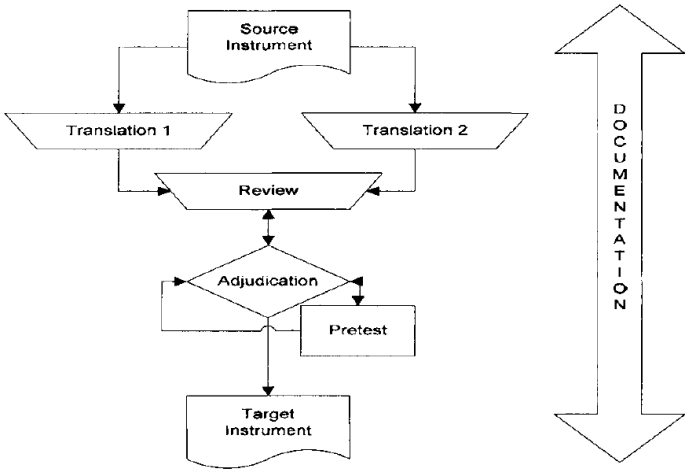
R: All items to be discussed in a review discussion

A: Adjudication of open queries, if needed

P: Qualitative and/or quantitative pretest of the entire questionnaire with a sample of the target population

D: Documentation

Figure 4.3. The TRAPD approach to translation.



67. The first step in the TRAPD approach is the production of two or more independent parallel translations (T in TRAPD). Relying on a single translation without any quality checks poses significant risks for subsequent fieldwork. The risks and thus disadvantage in using a single translation without any additional quality checks is that translation errors or poor wording, such as inappropriate style, or uncommon expressions or incorrect ways of addressing the target population, will be fielded without any quality checks. Even highly trained and experienced translators can make mistakes or use idiosyncratic wording or an inappropriate style if such translations are fielded without communication norms. Starting from at least two translations ensures translation options, reveals different interpretations, identifies idiosyncratic wording, uncovers obvious translation mistakes, and sparks discussions on challenging terms

like "household," "race," and "ethnicity." Even trained translators with similar backgrounds often produce slightly different translations of the same source text. [12]

68. The second step in the TRAPD approach is the discussion of all translated items in a Review discussion (R in TRAPD), which includes both translators and at least one social scientist or survey expert playing the role of reviewer. In the review discussion, the quality of the two initial translations is assessed and a new version of the target translation is produced. This can be one of the two initial translations, a combination of both translations, or a new translation developed during this expert discussion. Discussing every item of the source questionnaire in an in-depth expert review is at the heart of the TRAPD approach.

69. The third step of the TRAPD approach, Adjudication (A in TRAPD), comes into play when no agreement on a pre-final version can be found in the review discussion. It is led by the adjudicator, i.e., the person who makes the final decision on the translated questionnaire. The two roles, reviewer and adjudicator, are ideally covered by two separate people, but one person can serve in both roles if required due to financial and/or time constraints as a so-called “reviewer-cum-adjudicator”. The adjudicator should be a survey expert of senior standing with ample experience with surveys in the target language and country, and a decent level of fluency in both the source and the target languages.

70. The fourth step in the TRAPD approach is Pretesting (P in TRAPD). Pretesting translated questionnaires is just as crucial as pretesting monolingual questionnaires because it provides valuable feedback from a sample of the final target population. For translation purposes, quantitative pretests are useful for empirically testing the translated instrument and identifying any unusual response patterns that might be caused by translation errors or imprecise translations (Behr & Shishido, 2016). Qualitative pretests, such as cognitive interviews, focus groups, or debriefings with respondents or interviewers, offer deeper insights into how respondents understand specific terms or expressions. These qualitative pretest results can help confirm or correct translations (Schoua-Glusberg & Villar, 2014; Behr, 2017). See Section 4.2.1.5 for a more extensive discussion of questionnaire evaluation and testing approaches, all of which are applicable for assessment of both source language and target language questionnaires.

71. The fifth and final step in the TRAPD approach is Documentation (D in TRAPD). Documentation can serve various purposes and cover different aspects, and in particular lead to gains in efficiency of the survey process as well as to consistency of translations over time. First, during the process of TRAPD for a survey, it is beneficial to record translation problems encountered and decisions made at earlier stages, such as the Translation step. This way, if similar issues arise later, like at the Review step, the same discussions do not need to be repeated. In long-term survey projects where similar questionnaires are translated multiple times, documenting translation problems and decisions from one "wave," "round," or "edition" of the survey can similarly prevent redundant discussions and speed up the process if similar issues arise later. Essentially, it avoids reinventing the wheel. Additionally, documenting translation decisions helps ensure consistency over time, which is particularly important for repeated surveys.

[12] When working with a limited translation budget, a split approach can be used during the initial translation production. Instead of both translators covering the entire source questionnaire, each translates only half of the text. When using a split approach, alternate the distribution of the source text, similar to dealing cards in a card game. This ensures that both translators consider all parts of the questionnaire, even though they are only paid for translating half of the text (Harkness et al., 2010b).

72. Another important aspect to document is the chosen translation method, its implementation, and the individuals involved in the translation task. This includes their experience in (questionnaire) translation, background, qualifications, and current positions. While documenting these details may seem burdensome, it pays off in the long run. For example, if similar source questionnaire elements, such as entire questions, phrases, or answer categories, are repeated, previous discussions and decisions can streamline future

Box 4.6. The perils of back translation

While the method of back translation is still frequently applied, survey methods best practices strongly recommend against its continued use. From a Translation Studies point of view, there are different reasons why a simple forward and backwards translation - with the aim to assess the forward translation's quality (BTA - back translation assessment) – is not effective. First, if neither the forward nor the backwards translator have the chance to comment on their translations, many elements of a "good" translation will get lost: good translations are often no literal or word-by-word translations, but include the appropriate level of freedom in translation, because they need to express the meaning of the source text and not its words (Repke & Dorer, 2021).

When simply translating a forward translation back, it is very unlikely that a functionally equivalent, thus, a good translation would be carried out in both the forward and backwards translations. The highest level of "matches" between forward and backward translation will most likely be when both the forward and back translation are close or literal translations. But these are known to be of poorer quality because they are unlikely to capture the intended meaning of the source text but rather remain at the superficial level (Behr, 2017).

If forward and backwards translation both only rely on one person, there is the risk that these individuals produce erroneous or at least suboptimal translations (see the argument for parallel translations in the T step in TRAPD), which adds to the unreliability of back translation as method to detect translation errors.

Back translation has been found to be unreliable in detecting questionnaire translation problems. For example, Behr (2017) compared the errors detected by back translation to those detected by experts and found that "...While back translation can uncover problems, it causes quite a number of false alarms, and even more importantly, many problems remain hidden." (Behr 2017). In a recent experimental case study, Behr & Braun (2023) compared back translation to a team translation approach, in which only 24% of translation issues were addressed and solved by the back translation, whereas all of them were addressed and solved in the team translation setting.

The team translation approach has several advantages: having at least two independent translations provides alternative translation options and possibly different, idiosyncratic interpretations of the source text. The discussion by interdisciplinary experts, based on more than one initial translation, allows an in-depth work on the source text and joint efforts to develop the best-possible ways of expressing the source text's stimuli. The interdisciplinary discussion is crucial because all skills coalesce to contribute to the best-possible result: translation competence, linguistic skills in both the source and target language, and survey expertise in the target culture as well as topical understanding of the concepts to be implemented; all of these are required to develop questionnaire translations that will work to ask the intended questions.

translation tasks. Additionally, if translation errors are found, tracing back the team composition can help identify reasons for suboptimal translations and prevent them in the future.[13]

73. [[Case study requested on implementation of TRAPD in an LMIC.]]

4.5.4. Identifying suitable translation team members

74. Every method is only as effective as the people implementing it. When using the TRAPD method, it is recommended that at least one, preferably both, of the parallel translations be done by trained translators with academic qualifications in translating into the target language. If such training is unavailable for certain languages, seek professional translators with other relevant training. Additionally, all questionnaire translators should receive specific training and briefings on translating questionnaires, as this is a specialized text genre not typically covered in standard translation curricula. This should also include teaching some basics about surveys. Reviewers and adjudicators should be experienced survey experts with knowledge of questionnaire design in the target language and culture (European Social Survey, 2022). The translation team, consisting of at least three people (two translators and one reviewer/adjudicator), should discuss the translation of the questionnaire item by item and make informed decisions in the Review step.

4.5.5. Guidance and training for translation teams

75. Translators cannot work in a vacuum; they need context, background and specific instructions. We recommend conducting in-person training for everyone involved in the translation process; for trained and experienced professional translators, a training lasting 2-3 hours may be sufficient to teach them the basics of questionnaire translation, where questionnaire translation needs differ from other text genres, etc. If necessary, training can also be virtual, depending on financial and geographical constraints. Experience shows that direct interaction during training and briefings is more effective than relying solely on written or recorded guidance materials. Guidance for translation teams is needed at two levels: (a) instructions about the translation method, and (b) translation instructions at survey question level.

4.5.5.1. Translation guidance on the general translation method

76. The translation method to be implemented should be clearly outlined in a central guidance document, often referred to as "Translation Guidelines" or the "Translation Manual." This document is typically drafted by a central team and made available to everyone responsible for questionnaire translations, including those in participating countries, agencies organizing or conducting translations, and the translators themselves.[14]

77. This central guidance document should cover all elements of the translation approach. It should clearly explain the translation method (ideally TRAPD) and the process to be followed, including details about the translation platform. It should outline the requirements for participants in the translation process, clearly define each participant's roles, and provide contact information for process and translation queries. Additionally, it should provide briefing information about the overall survey, including details about the target population, study background, and the expected difficulty level of the instrument.

78. Even experienced and professional translators need specific instructions because translating questionnaires is rarely covered in standard translation curricula. Without proper guidance, translators

[13] See Behr et al., 2019 and Behr & Zabal, 2020, for more on the importance of documentation for questionnaire translations.

[14] See the Translation Guidelines for the European Social Survey (European Social Survey, 2022) and the current translation protocol for USAID's Feed the Future surveys (ICF, 2024) for examples of translation protocols.

might choose an inappropriate language register. For example, they might select a language level that is too high, resulting in translations that are elegant but not easily understood by less educated respondents in the target population. The wording of particular elements of questionnaires, such as answer categories, will also often need to follow certain questionnaire design rules that are not intuitive to linguists without knowledge about questionnaires (e.g., the need to be consistent across the questionnaire, to carefully differentiate between quantifiers in fully labelled scales, or to translate extreme anchors such as “fully, completely, extremely” in an extreme sense). Knowing that a simpler, more understandable phrase meeting questionnaire design requirements (“survey speak”) is preferred will help create a translated instrument suitable for fieldwork purposes.

4.5.5.2. Translation guidance at the level of individual questions

79. Translation guidance at the level of survey questions should provide specific instructions on translating certain parts of the source questionnaire. This guidance may be referred to as translator notes, footnotes, annotations, or item-by-item guidelines, depending on the project. These notes can be added directly to the source questionnaire or listed in a separate file.

80. Even if the source questionnaire was developed with future translation in mind, methods like Advance Translation, as discussed above in Section 4.2.8, can help the central team anticipate certain source language words or expressions to pose translation challenges. For example, "polyseme" words have multiple meanings, and it must be clearly defined which meaning should be conveyed in the translations. An example of a polyseme word in English is "fair," which can mean "marked by impartiality and honesty" or "not very good or very bad," depending on the context. Item-level translation guidance provides instructions on how to translate or not translate specific parts of the source text, such as individual words or expressions. This guidance helps harmonise translations across languages, and minimize translation problems, and the risk of errors due to difficult or ambiguous source text.

81. For “trend” items that are repeated within the same survey project, the same translations, without any change, should be reused. Maintaining identical wording over time contributes to measuring true change between surveys. However, when a project re-uses existing items from other survey projects, the project must decide whether they trust the translation quality level of the other survey, or if all translations of source language questions are to be carried out following the organization’s translation process, ideally TRAPD.

4.5.6. Gender in questionnaire translation

82. Gender is a significant consideration in questionnaire translation at linguistic, cultural, and political levels. Linguistically, different languages handle gender in various ways:

- **Gendered languages** have grammatical gender (e.g., Slavic or Romance languages).
- **Languages with natural gender** (e.g., English) use gender-specific pronouns.
- **Genderless languages** (e.g., Finnish or Hungarian) do not have grammatical gender.
- **Gender-neutral languages** (e.g., Thai, Vietnamese, Chinese, Philippine languages)

83. For languages with grammatical or natural gender, gender issues are more prominent, requiring careful decisions when drafting questions in the source language and translating them into the target language (Prewitt-Freilino et al., 2012).

84. In questionnaires, gender considerations are important at several levels:

- **Respondent gender**: When addressing the respondent (e.g., "Can you please tell us …?"), the text needs to be adapted to the respondent's gender. This includes gendered adjectives in answer categories. For example, in French, "very satisfied" would be "satisfait" for male respondents and "satisfaite" for female respondents.
- **Interviewer gender**: In some languages, phrases like "I am now asking you …" need to be adapted to the interviewer's gender.
- **Gender of other parties mentioned**: For example, "How much do you trust teachers?" may need to specify whether it refers to male teachers, female teachers, or teachers of any gender.

85. Decisions on handling gender in questionnaires will vary by country and often require balancing different considerations, such as whether to mention all genders or exclude one systematically. An exception might be if the source language intentionally asks about only female or male doctors. Additionally, in gendered languages, it must be decided whether answer categories with adjectives should vary based on the respondent's gender, where English uses a single form (e.g., "very satisfied" in English vs. "très satisfait.e" in French). Translators and team members need to decide if this should be done consistently across the questionnaire and, if so, how to format it (e.g., using a dot, asterisk, brackets, or spelling out both words). Another consideration is the more recent inclusion of a "third gender," which does not follow binary gender differentiation. Some languages have grammatical solutions for this, such as the English use of "they/their" to include any gender, the French pronoun "iel," or the Polish "Państwo." However, these new forms may not yet be widely accepted in some countries and if incorporated into a translation, should be carefully pretested for sensitivity and social desirability bias.

86. When asking about the respondent's gender as a background variable, it is important to consider whether to offer a third gender option and, if so, in what form. Terms like "normal" or "other" can be discriminatory, so it is crucial to choose wording that respects everyone's gender identity. There is extensive research on how to ask about gender in cross-cultural surveys (e.g., Hadler et al., 2022).

87. Gender considerations are especially important in self-completion modes. While interviewers can adapt to the respondent's gender during an interview, all decisions must be made beforehand in self-completion settings. In web-based surveys, different gender versions can be programmed. However, in paper-based surveys, all questionnaires must be printed in advance. This raises the question of whether to create a single, longer version that includes all genders or to print different versions for each gender. Providing all gender forms across the whole questionnaire may trigger a higher cognitive load for the respondents, as the text becomes longer. Both the decision to include and to exclude all gender forms risks alienating certain respondents and non-response bias may occur regardless of the approach. No recommendation can be given centrally that would be applicable for any questionnaire translation. The decisions have to be taken by those responsible at project level, taking the above points into consideration, and guided by what is most common in each country at the time a survey is to be fielded.

4.6. Questionnaire evaluation and testing

88. As Willis (2019) emphasizes, evaluation and testing of questionnaires are crucial within the QDET framework for identifying and reducing measurement error and ensuring the accuracy and reliability of the data collected. Questionnaire evaluation refers to expert review of the questionnaire using different frameworks. Testing refers to administering the questionnaire (in different ways) to "real world" members of the target population and recording feedback. These evaluation and testing activities all help researchers assess both individual survey questions as well as the overall flow of the questionnaire before data collection begins. Importantly, testing can also identify specific questions as sensitive in particular contexts and

consider revisions to either reduce sensitivity or increase respondent confidence in the confidentiality of their survey data.

89. **Table 4.2** provides an overview of the most common questionnaire evaluation and testing approaches, along with an explanation of how best to incorporate each into the questionnaire design process. These techniques are often used after developing and translating survey materials, but some, like focus groups as discussed in Section 4.2.1.1, are used earlier to inform research design aspects such as question wording and the target population.

Table 4.2. Questionnaire evaluation and testing methods.

Evaluation Methods			
Method	**Description**	**Strengths**	**Weaknesses**
Expert review (Groves, et al. (2009; Tourangeau et al., 2019)	Review of draft materials by experienced methodologists	Cost efficient and quick; can identify a wide variety of problems in the questionnaire; requires very small sample of experts (usually 2-3); nearly always worth doing (Tourangeau et al., 2019)	Subjective; variability in expert assessments (Ikart, 2019; DeMaio & Landreth, 2003); no respondents involved
Questionnaire appraisal system (QAS) (Willis & Lessler, 1999; Tourangeau et al., 2019; Dean et al., 2007; Schaad et al., 2020)	A systematic tool for evaluating survey questions, identifying up to 26 issues for each survey question, including potential problems with translation	Cost efficient; can be used to identify questions that would benefit from cognitive testing	Identifies issues in nearly all questions
Question Understanding AID (QUAID) (Graesser et al., 2006; Tourangeau et al., 2019)	Software that identifies technical features of questions that may cause comprehension problems	Good at finding problems with the syntax of the questions; inexpensive	Limited evidence that the results are more useful than those from other evaluation methods, like expert review
Survey Quality Predictor (SQP) (Saris et al., 2004)	A tool based on meta-analyses of Multi-trait-multi-method (MTMM) studies that provide estimates of reliability and validity for existing survey items	Low-cost; does not require collection of data	Requires subjective coding of questions by researchers before use, leading to coding errors; created using primarily attitudinal survey data

Continued on next page…

Table 4.2. *continued*

Testing Methods			
Method	**Description**	**Strengths**	**Weaknesses**
In-depth interviews and focus groups (Davis & DeMaio (1993) for an overview; also Groves, et al. (2009))	One person (in-depth interview) or small group of people (focus group) brought together to discuss specific topics, led by a moderator	Useful when there is limited information on the topic of interest; uses the same types of respondents who are the target population for the survey; allows for immediate follow up; requires small focus group size (10-12 participants) or individual participants (for in-depth interviews)	Mainly qualitative; results should be carefully interpreted due to small sample size; requires well trained moderators; small group dynamics in focus groups may influence the results
Cognitive interviewing (Beatty & Willis, 2007; Willis, 2005)	In-depth interviews with a small sample of respondents (5–12 per target language) who answer survey questions and then administered follow-up probes to target the cognitive response process	Gathers data from the target population; reveal comprehension or recall issues that other methods would not identify; can assess how well the intended constructs capture comparability across diverse cultural groups (Miller, 2019);	Coding and subsequent systematic analyses is challenging; cognitive interviewing practices are not standardized; can be difficult to assess how common problems are in the target population
Field pretest and interviewer debriefing (Goerman, et al., 2007; Groves, et al. 2009; World Bank, 2024)	A pretest of the questionnaire in the field with a small number of respondents (generally at least 60 respondents) to test the survey instrument, ideally in all target languages, followed by a group discussion with interviewers to discuss experiences	Realistic; allows for testing of the questionnaire flow; allows for feedback from interviewers and adequate data for basic analysis to identify issues with skip logic and other survey instrument errors. Debriefing benefits from interviewers' experience about what made a question difficult and with what types of respondents	Financial cost; requires larger sample size relative to the other techniques, requires logistical coordination; respondents may answer questions even if there are comprehension problems, without the interviewer or the researcher being aware of the issue; interviewers themselves may be partially responsible for respondent confusion with a question
Usability Testing (Hansen & Couper, (2004; Tarnai & Moore, 2004; Geisen & Murphy, 2019; Nichols et al., 2019)	Testing of the functionalities of the survey instrument to assess the electronic data capture functionality, or testing of paper materials; see also Section 4.8 of this chapter	Direct user assessment of the tools that will be used during data collection; requires few resources as employees of the survey organization can be involved in testing; usually requires small sample sizes	Time consuming; respondents are generally not involved

90. Other, less-common methods of questionnaire testing include:

- Behaviour coding, which involves systematic coding of the interviewer-respondent interaction in order to identify problems that arise during the question-answer process (Mangione et al., 1992; Groves et al., 2009)
- Vignettes, which are brief stories or scenarios describing hypothetical situations or people to which respondents are asked to react, allowing the researcher to explore contextual influences on the respondent's cognitive response process (see Rossi & Anderson, 1982)
- Item Response Theory (IRT), which are statistical models that facilitate examination of the ways in which different items discriminate across respondents with the same value on a trait (see Reeve & Mâsse (2004))
- Latent Class Analysis (LCA), which are statistical models that allow examination of error rates associated with different items (Yan et al., 2012; Kreuter et al., 2008).

91. [[Case study requested on the use of cognitive interviewing for questionnaire pretesting in an LMIC.]]

92. An effective pretest design combines various pretesting techniques to leverage their strengths and minimize their weaknesses. These methods often complement each other, enhancing the overall efficiency of the pretest process. For instance, starting with an expert review can help refine the questionnaire based on expert feedback, followed by cognitive interviews or a pilot study. However, studies have shown that different pretesting methods can yield varying and sometimes contradictory results regarding survey question performance (Rothgeb et al., 2001; Forsyth et al., 2004; DeMaio & Landreth, 2004; Jansen & Hak, 2005; Beatty & Willis, 2007; Yan et al., 2012). Therefore, it is crucial to carefully consider the strengths and weaknesses of each method, as well as factors like timeline, available resources, language, cultural norms, and interviewer characteristics, when selecting pretesting techniques. The most suitable combinations of methods may vary across different countries, and this variability should be considered when evaluating and comparing pretest results.

4.7. Determine data collection systems

93. As noted earlier in this chapter, most surveys use an electronic data capture system to collect data using a **survey instrument**, whether the mode is face-to-face surveys (CAPI), phone (CATI), or self-administered (web). The survey instrument has multiple components. The most visible is the programmed questionnaire. However, the instrument also contains other critical components, including sample management, question filtering and skip logic, randomization of question order and response options, management of experimental question designs, data quality monitoring tools (e.g., through a quality-monitoring dashboard), autodialers (CATI), and contact systems (web). There are a variety of electronic data capture options, and systems are continuously evolving in functionality. This section provides an overview of the components to consider when determining an appropriate data collection system, while the next section (Section 4.8) reviews best practices for programming and testing. Section 4.4 provides a list of several electronic data capture software systems more commonly used at the time of the publication of this Handbook.

94. It is important to carefully evaluate the software options available to ensure the system will meet the specific needs of the survey (and the stakeholders, as relevant). Ideally, this decision will be made at the organizational or project team level, taking into account all types of surveys they are likely to field and selecting software that will meet as many anticipated needs as possible from its typical stakeholders (e.g.,

randomization of modules, geofencing, etc.), limiting the number of software platforms the team will need to learn and finance.

95. While many software platforms are available (see Section 4.10 for a list of some commonly used software options), each has specific capabilities and limitations. Organizations must understand and define the needs of the data collection before evaluating and selecting an appropriate software; changing software solutions later can be costly and time consuming. Software options are constantly evolving; new platforms are introduced and new functions are regularly added to existing platforms. As a result, this chapter provides a framework for evaluating options and understanding of best practices to employ when using electronic data collection software and systems. There is no single, best software option; the best option is one that effectively meets the needs of data collection projects your research team is implementing.

4.7.1. Hardware compatibility

96. Once a decision is made on the software system that will be used, hardware must be selected and acquired that meets the needs of the software system.

- The most common devices used for CAPI data collection are smartphones, phablets (smartphones with large screens), tablet computers, or laptops. Some software systems are compatible with multiple devices while others are limited to specific platforms, such as Android mobile or Windows. In addition to the operating system on the device, software systems often have specific requirements for processor speed, memory, and internet connections that must be met by the hardware being used. For surveys that contain visual stimuli that must be shown to a respondent during the interview, smartphone screens could limit the size of the image compared to what is possible with phablets and tablets and may make it more difficult for respondents to easily view the stimuli.

- Web based surveys (CAWI), where a respondent self-completes the interview, used to assume respondents will use a laptop or desktop computer, but should now be designed to allow for completion on smartphones and tablets as well because these smaller, personal devices are rapidly becoming the most common devices respondents use to complete surveys. Researchers must account for a wide variety of possible devices (screen sizes, makes, and models) and select software that allows for as many options as possible to optimize user experience and, ultimately, data quality (Toninelli & Revilla, 2019).

- For CATI data collection, it is essential that the selected software will work with devices that can be acquired by the organization by purchase, rental, or shipping to the location where data will be collected. Most CATI systems use laptop or desktop computers within call centres, though systems are now available that work through an application on a smartphone or tablet, allowing interviewers to work outside of a traditional call centre environment.

97. While it is generally not possible to provide visual stimuli to respondents for CATI data surveys and those designs should focus on interviewer readability, the visual display of survey items can carry additional importance for CAPI and CAWI surveys where respondents must interact directly with the screen and page design. Determining the visual display requirements of the survey will inform the decision on device screen size and should be determined before selecting the software package.

- The research team must be able to produce digital versions of stimuli and the software platform must be able to handle the import of those digital images.

- Survey questions that rely on complex grids may not be able to be easily displayed on smaller screen devices; tablets may be better than smartphones in this instance.
- Navigation controls (moving through questions and screens) should be clearly visible and intuitively usable for both self and interviewer administered designs.

98. Each software platform has a minimum list of technical specifications that devices need to meet to properly run the software and these should be well understood, verifying that the devices meet the minimum specifications for the software platform. Not all devices are available in all regions or countries so it is also important to verify availability of devices before making a decision on the software as certain platforms may not be able to be used on the devices available to the research team in the location of the data collection.

4.7.2. Connectivity requirements

99. It is generally preferred, and often a necessity, to select survey software that is fully functional without a constant internet connection. Field teams often need to conduct interviews in remote or rural areas that are not within range of a telecom or Wi-Fi network so the software should remain fully functional when no internet connection is available.

100. Collected data will be stored locally on the interviewer's device until they are able to connect to the internet and upload the completed surveys. Uploads of data can be completed in a variety of ways, including (1) central servers located at the data collection team's home office, (2) cloud servers that host data, or (3) a proxy server such as a tablet or laptop operated by a field supervisor. The research team must first determine the type of internet connection(s) available to the team throughout fieldwork and ensure that the internet connectivity available will work with the type of upload solution offered by the software platform.

101. If field teams will use a telecom network to upload data, then the data collection devices either need to have the capability to use SIM cards or the field team will need to be provided with a SIM card enabled hotspot from which several field team members can connect and upload data.

102. If the field teams will rely on Wi-Fi connections to upload data, the tablets will need to have Wi-Fi capability and the field teams will need to have regular access to Wi-Fi connections throughout the field period. Data should be uploaded from the tablets daily, or as frequently as possible given the internet infrastructure and availability in the location of the data collection. Some data collection software allows for uploads to a proxy server (often a field supervisor's tablet or laptop), which may be the best solution when teams will be in remote locations without an internet connection for a long period of time because this will create a copy of the data that can be reviewed by supervisors during fieldwork and ensures that a backup of the data is maintained when uploads to a central server or cloud server can only be done sporadically.

4.7.3. Security features

103. Keeping data safe and secure is a key requirement of any data collection software. While no software system is completely infallible in this regard, the research team should verify that the software platform has safeguards in place to mitigate the risk of data leakage or loss. Systems are constantly upgraded to improve data security, but some core qualities should be present in any software used for collecting data and protecting personally identifiable information (PII).

- **Encryption:** data should always be collected and stored on encrypted devices and this should be a standard requirement for any devices used in the field. Data is most vulnerable when it is being transferred from the data collection device to the server used to store aggregated data from the survey project. To mitigate the risk of confidential data being intercepted during upload, the

software should use end-to-end encryption during any transfers over the internet. Using this process mitigates the risk that sensitive data could be intercepted and viewed by people outside of the research team by only transferring encrypted files during upload.

- **Remote wiping:** in the event that a data collection device is lost or stolen, the data collection software (or additional software loaded onto the tablet) should have the ability to remotely wipe all data from a tablet that is no longer in possession of the field team. While this would still lead to data loss, it will mitigate the risk of data leakage and the release of PII to unauthorized actors.
- **Passwords:** both the device and the software should be password protected to limit the risk of data leakage if a tablet is lost or stolen. Each interviewer and each data collection device should have a unique password that must be entered each time the device is opened and the data collection application is started.
- **Data transfer functions:** many devices contain multiple modes of sharing data between devices. Some software platforms may take advantage of these functions by allowing cases to be transferred between devices via Bluetooth, for example. While this type of functionality can be useful, if these functions are not being used for the data collection, it is safer to disable those functions on the tablets to mitigate the risk of data leakage. For example, if data will only be transferred and uploaded via WiFi connections, disabling Bluetooth and SIM card connections is best practice as this limits a potential source of data leakage or device hacking.

4.7.4. Sample management system

104. Sample management is generally driven by the mode of sample generation (compiled lists of respondents, random digit dialing (RDD), random selection of households in the field, pre-selection of households, etc.). In some cases, the software will need to have the capability to upload information about the sampled respondents before field work begins while in other cases, the software must be able to support the process being used for randomly or systemically sampling respondents during field work. In both cases, the software must be able to record the results of each attempted contact and compile those results for later analysis.

105. Each software system handles sample management differently and the research team must carefully define their requirements for sample management to ensure the software will handle those requirements. The team should outline the full sampling process and verify that the software they select will handle all of the steps of sampling for the project.

- **Unit/household selection (CAPI and some CATI projects): I**f there will be a listing of all units/households in an enumeration area prior to selection, some software platforms allow the field teams to conduct the listing on the survey platform and upload results of the listing to a proxy server or remote server so that units/households can be selected, either randomly or purposefully, prior to data collection. For some projects, an additional sampling step within a selected unit/household will need to be made (one member within the unit/household must be selected). Some platforms allow these steps to be done completely in the field without the need to upload listing data to a central server while others require an internet connection to enable this function. As noted above, understanding the internet infrastructure in the location of the data collection will help guide decision making on this requirement.
- **Pre-selected units/households (all modes):** Some software allows for files of preselected sampling units to be sent to auto-dialers, central servers, or directly to data collection devices so

that the sample can be used to make initial contact with a respondent. For CAPI surveys, interviewers will have the selected units they are responsible for interviewing loaded onto their devices prior to data collection. For CATI and CAWI surveys, the dialer or server will have the selected units stored centrally. In all of these cases, this could be listed sample (names, contact information, and locations of pre-selected respondents) or randomly selected sample (using an RDD generation method or using GIS selection where GPS coordinates of selected units are sent to the devices and interviewers can view these on maps in their devices to determine which units/households should be interviewed).

- **Random selection in field (CAPI):** If a random walk or other similar methodology is being employed to select units/households for an interview, the software should be programmed to guide the interviewer through the full process. This includes instructions for the initial sampling of units/households in addition to any screening protocols or questions that may be included as part of the selection of respondents within selected units/households (head of household, most knowledgeable on a certain topic, most recent birthday, kish grid, etc.).

- **Respondent selection within units/households:** When randomizing selection of a respondent within a household, most software options will accommodate a screening process such as "most recent birthday" by simply programming that question into the survey instrument. Most software platforms can also accommodate randomization of respondents within households by using a roster function that lists all eligible household members and then the software will randomly select one respondent for the interview. Many platforms also allow the research team to program more complicated rules for selection, based on data entered into the household roster if the survey requires a more complicated respondent selection process within the unit/household.

- **Disposition coding:** Many software platforms allow for the recording of attempts directly into the software platform. These systems allow interviewers to record refusals, failed screeners, call backs, and other dispositions into the same system used for administration of surveys. Many software platforms allow interviewers to record temporary dispositions for attempted households or sampled units (call backs and other situations where the interviewer will return to make another attempt) in addition to final dispositions (completions, refusals, non-target, language barrier, etc.) for households or units that will not be revisited (see Chapter 6 on coding dispositions during data collection). Each software platform has its own way of handling disposition coding with differing methods of resuming cases once they have been opened. It is important to understand both the benefits and the limitations the software platform offers in this area and to properly train field teams on how to use these functions in the field to ensure accurate and complete documentation of household attempts are manageable within the software platform.

106. Some software platforms do not manage sample or case dispositions in a manner that works for data collection but contain other functions for administering the survey that are essential to the project. In such cases, it may be necessary to use a more traditional pen and paper method for sample management or case disposition documentation and manually manage those processes throughout field work. For example, the Tangerine software platform does not manage sample or have a simple means of recording case dispositions, but it has specific designs used for reading assessments that are hard to replicate in other software platforms. In cases such as this, using pen and paper or using another software platform that manages sample and case dispositions may be necessary. These approaches must be carefully thought through as they will require additional steps to connect sampling and disposition data back to survey data after field work is complete. With additional complexity comes the risk that these linkages will be lost

during analysis or that data will be lost through interviewer error or oversight while navigating multiple systems simultaneously.

4.7.5. Cost

107. While there are many survey software platforms that are open source and free to use, some platforms require either the purchase or rental of licenses to use the software. Some platforms charge for the application while other platforms only charge for the use of server space. Costs often vary based on the size of the survey project or charge monthly rates for the use of the software or server space. It's important to carefully review the terms of cost agreements prior to committing to a platform and ensure that all software costs are included in the project budget ahead of time. It is also important to fully understand the needs of the survey when evaluating the cost of the software system to be sure that there is not a free or more cost-effective option that meets the requirements of the survey rather than a costly option that has more functions but functions that will not be used for any particular survey. Just because a software platform has more functions does not necessarily mean it is a better match for your project.

4.7.6. Programming

108. Software platforms only work when they are set up effectively and correctly. There is a wide range of complexity in programming within different survey platforms. Some are designed to be very user friendly so that most people with basic computing skills can learn how to program within the platform quickly. Other platforms require complex code to be written in order to program the instrument, in which case a consultant who is highly skilled and experienced may need to be engaged to handle the programming of the instrument into the platform language. If a complex platform is selected and the research team does not have the experience to program the instrument themselves, the team should include the cost for an external consultant in their budget to support programming.

4.7.7. Multi-mode systems

109. If the survey will be fielded using different modes, some software platforms have the ability to manage the same survey through multiple modes natively. For example, if a survey will be fielded both in person on computer devices and online through the web, some software platforms can be used to field the same survey through both modes. This may improve efficiency for the research team as it can eliminate the need to process and merge data from multiple platforms and instead house completed data from multiple modes within the same system. Some platforms can manage CAPI, CATI, CAWI, or SMS or various combinations of those modes. Defining all modes that data needs to be collected through can help guide the platform selection decision.

4.8. Programming and testing

110. This section focuses on programming the questionnaire specifications (see Section 4.4), including testing and issue tracking of the programmed instrument.

4.8.1. Programming the questionnaire specifications

111. Programming and testing a survey instrument are critical phases in questionnaire development, ensuring that the questionnaire functions correctly and efficiently during data collection (Fowler, 2014). To produce a high-quality and effective survey instrument, a survey programmer should program a survey instrument with three user groups in mind: the interviewer, in the case of interviewer-administered surveys; the respondent who provides the information; and the analyst who will analyze the data (Gibson & Louw, ND). During the programming stage, it is crucial for programmers to adhere to the instructions outlined in the survey specifications. All question text, response options, and routing should be determined by the

questionnaire designers and stakeholders and documented in the specifications. Programmers should not make any modifications to these elements without prior consultation and approval. **Table 4.3** provides programming recommendations to optimize functionality, based on the perspectives of three users: interviewers, respondents, and analysts.

Table 4.3. Programming recommendations to optimize functionality.

User	Programming recommendations to optimize functionality for the user	Example
Interviewer	Format the text as needed to emphasize words and distinguish between respondent-facing text and interviewer instructions.	Consider displaying respondent-facing text in black font and interviewer instructions in blue font.
	Let the interviewer know which text should not be read to the respondent.	Interviewers are typically trained to read all response options, unless noted otherwise. In most instances, response options for "Don't know" or "Prefer not to answer" should not be read. This should be indicated by the use of "(Do not read)" after each of these response options
Respondent	Program fills into questions that rely on previous responses to reduce the time spent looking back at previous responses.	A questionnaire module asks questions about a randomly selected child in the household. Make use of fills to auto-populate the child's name into the questions for this section.
	Minimize the amount of time and effort to complete the questionnaire by having the programmed instrument perform complicated calculations for the respondent.	Calculate total hours worked based on average hours worked per day and number of days worked per week.
Data analyst	Use self-explanatory variable names so that it is easy to find variables when doing analysis.	Consider a combination of letters and numbers that are short and intuitive for data analysts. For instance, demo_age could be a variable name for a question on age in the demographics section of a questionnaire.
	Use conventions for value labeling or naming for answers that are common to multiple questions.	Consider assigning the values of "-88" for Other, "-98" for Don't know, and "-99" for "Prefer not to answer". Similarly, assign "0" for No and "1" for Yes.
	Program soft and hard constraints to check, in real-time, for inconsistent or outlier responses.	Hard constraint (must be satisfied to proceed in the questionnaire): Respondent cannot be older than 120 years. Soft constraint (can be violated): A respondent's income is above a predefined value.

112. To minimize the number of programming errors, we suggest a phased approach to programming.

- **Phase 1:** Begin by programming a "skeleton questionnaire," which includes only the programming of essential elements: variable names, question text, response options, and interviewer instructions. This skeleton serves as the foundation for the survey instrument and provides a framework for subsequent programming steps.
- **Phase 2:** Once the skeleton questionnaire is programmed, it is crucial to test its completeness and accuracy. Testing at this phase ensures that all questions, response options, and instructions are correctly implemented and that the basic structure of the survey is sound. We recommend unit testing (see below) to test the accuracy of the skeleton questionnaire.
- **Phase 3:** After confirming the accuracy of the skeleton questionnaire, advanced components of the instrument can be programmed. This includes incorporating translations and survey logic, such as skip logic, enabling conditions, and randomizations, which are necessary for navigating the questionnaire correctly and ensuring that respondents are presented with relevant questions.

4.8.2. Testing the questionnaire specifications

113. Once programmed, a survey instrument should be thoroughly tested. There will likely be several iterations of programming and testing before the programmed instrument is finalized. Developing a comprehensive testing plan is essential to guide the testing process and ensure that the instrument is thoroughly evaluated. This plan should outline the testing timelines and roles and responsibilities, as well as describe in detail the methods for testing the survey instrument. Some examples of testing methods are described below.[15]

- **Unit testing:** Compare survey specifications to the programmed instrument to ensure that questions, response options, and enabling conditions are correctly programmed (Hansen, et al., 2016; Myers & Well, 2003). The purpose of this method is to identify discrepancies between the questionnaire specifications and the programmed version. This testing method can be carried out by the survey programmer or staff not involved in the programming of the questionnaire.
- **Testing media question**: In some cases, survey instruments may require interviewers to capture audio recordings or photographs using their phone or tablet cameras. During the testing process, it is important to assess image and audio quality to ensure the responses captured in these questions meet the required standards. Additional checks should include verifying that the media files are appropriately sized and can be easily transferred via cellular or Wi-Fi connections, or otherwise retained on interviewer devices until the files can be manually transferred when interviewers return to the office at the end of data collection.
- **Scenario testing:** Create predetermined scenarios (responses to questions) to ensure that the programmed instrument's content and logic accurately reflect the questionnaire specifications. In some instances, it may not be possible to test every conceivable path in the questionnaire, however, predetermined scenarios will verify that the programmed survey behaves as expected. When developing scenarios, it is important to thoroughly review the questionnaire specifications to identify all possible pathways that need to be tested. Once the scenarios are drafted, someone

[15] See also How to pretest and pilot a survey questionnaire - tools4dev

else should review prior to testing to ensure accuracy and completeness. Testing should then be be conducted by staff not involved in the programming of the questionnaire or by interviewers.

- **System testing:** Conduct at least 20 test interviews and export the test data. The Data Processing team can review this test data to verify that the instrument is programmed correctly and that the data collected is accurately exported to the statistical software as intended. This method will generate cross-tabulations to check skip logic and routing, ensuring that the survey logic is functioning as intended. One of the benefits of the system testing approach is that it allows testers to conduct test interviews as if the survey were live. This testing method should be carried out by staff not involved in the programming of the questionnaire or by interviewers.

- **Dummy data:** Generate dummy data (faux data) and export to statistical software for further review (note that this feature is not available with all data collection software). Dummy data provides an opportunity for a thorough examination of the instrument's programming and data output, particularly, variable names and types, response options, and programming logic. This testing method should be carried out by the survey programmer or data analyst (for Qualitrics: https://www.qualtrics.com/support/survey-platform/survey-module/survey-tools/generating-test-responses/; For SurveyToGo: https://support.dooblo.net/hc/en-us/articles/208295305-Using-the-SurveyToGo-Dummy-Data-Generator).

- **Pretest:** Conduct a small-scale survey pretest that simulates the data collection process for the study (see Section 4.1.1 for more information about Questionnaire Design, Development, Evaluation, and Testing (QDET) framework). The pretest is an opportunity to gather feedback on issues or challenges encountered in questionnaire administration (by interviewers) and questionnaire comprehension (by respondents), including errors in instrument design, programming, and translations (https://dimewiki.worldbank.org/Survey_Pilot). See the section on pretesting above (4.6) for additional details.

114. In addition to a testing plan, it is essential to make use of a tool for tracking the status of issues flagged during the testing period. Effective issue tracking ensures that problems are promptly logged, addressed and resolved. There are many tools to keep track of issues, and the choice of tool will vary between projects and teams, as well as informed by the data collection platform used to administer the questionnaire. Issue tracking can be done manually with Microsoft Excel or through online tools such as Jira, Sifter, Microsoft Lists, Zoho, Microsoft Azure DevOps. At a minimum, the issue-tracking tool should include fields for the following information: the creation date, a description of the issue, the type of issue (such as question wording, response options, translations, programming logic, or interviewer instructions), the severity of the issue, a description of potential solutions, the current status, the person assigned to resolve the issue, actions taken to address the issue, and the last change date.

115. The lead survey programmer plays a crucial role in monitoring the issue tracker, regularly reviewing the logged issues, entering and responding to comments, and making necessary changes to the programmed instrument and questionnaire specifications. By actively managing the issue tracker, the lead programmer ensures that all identified issues are systematically addressed, minimizing the risk of errors during data collection. This proactive approach not only improves the reliability of the survey instrument but also facilitates a smoother transition from testing to full-scale data collection, ultimately contributing to the success of the survey project. Lastly, careful data quality monitoring can detect issues in the programming of the questionnaire even after fieldwork begins, allowing for revisions to the program and deployment of updates to the interviewers during fieldwork.

4.9. Emerging approaches

116. Section 4.9 provides an overview of several emerging approaches relevant to the field of questionnaire design at the time of the publication of this Handbook. In particular, the rapid development of machine learning and artificial intelligence (AI) technologies in recent years has brought new opportunities to the field of survey questionnaire design, translation, and in the areas of evaluation and testing of survey items.

4.9.1. Machine translation

117. Machine translation (MT), sometimes referred to as automated translation, refers to the use of a website or other tool that allows the user to copy-paste source language text and automatically produce a target language translation.[16] In MT, large bilingual text datasets ("corpora") are used to train "engines", which in combination with algorithms, provide rapid translations of the source text entered. The quality of MT differs between language pairs, with more common language pairs, such as English-Spanish, generally yielding better MT results than less common pairs, such as English-Latvian, where the engines have less text material to learn from. The quality of MT also depends on the type of engines used to produce the translations. For example, there was a significant improvement in quality when the MT engines moved from statistical MT to artificial neural-networks-based engines. Machine learning in combination with increased computational power has led to a marked improvement in MT quality in the early 2020s (Zavala-Rojas et al, 2024), and MT is more frequently used for translations.

118. Despite its widespread and growing use, MT has many weaknesses. Sometimes MT results are clearly wrong; for example, due to a misunderstanding of the context. For example, MT will have problems producing a translation of the agree-disagree scale into German that can be fielded as it may translate the answer categories (i.e., agree, disagree, neither agree nor disagree) into infinitive forms of these verbs, resulting in a nonsensical translation. Often, translations produced by MT seem, when looking at them superficially, to be correct and read nicely, but when translation experts assess the translation quality, they tend to discover serious omissions, errors or weaknesses that would be harmful if used for fieldwork and are easily undetected if human experts are not included in a translation team (Zavala-Rojas et al, 2024). Another weakness of MT is that the quality differs between language pairs (see above), with little quality control related to translation decisions by MT engines. Lastly, not all languages are available for MT, and important variations in regional dialects will also not be incorporated when MT is used. Therefore, the recommended approach is not to use machine-translated text in its raw form, but rather to apply "machine translation and post-editing (MTPE)", which means that a human translator reviews the MT raw output and edits it according to specific guidance.

4.9.2. Generative AI for questionnaire evaluation and translation

119. Generative AI tools such as ChatGPT have the potential to be used to evaluate survey questions, in an approach similar to a survey team member using the SQP tool to evaluate questions on a number of different metrics. Preliminary research indicates there is an opportunity for such AI tools to play a role in evaluating survey items and identifying issues in the questionnaire, and we expect that AI tools will continue to grow in sophistication and availability in the coming years (Olivos & Liu, 2024). However, as with other recent applications of AI discussed in this chapter, caution and careful review of AI-generated results is warranted, and users should consider the origins of training data for generative AI and the generalizability to the context to which generative AI is to be applied for any survey.

[16] Current freely-available tools include Google Translate and DeepL.

120. Generative AI has also been used in the context of questionnaire translation, with several use cases of ChatGPT involved in questionnaire translation. In a recent example, ChatGPT was used for assessing translation problems in TRAPD (Metheney & Yehle, 2024). At the time of publication of this Handbook, it is still too early to draft recommendations on the use of generative AI for the field of questionnaire translation; scientific evidence is needed. However, innovative and promising developments are ongoing.

4.9.3. Survey workflow tools

121. With the speed of technical developments, we expect, and partially already observe, rapid IT developments in the field of managing the survey lifecycle, bringing welcome efficiency particularly to the questionnaire design and translation of questionnaires for large-scale surveys of households and individuals. A successful example of an IT development to facilitate and streamline the survey lifecycle is the "DataCTRL survey tool suite", developed by Centerdata (Netherlands). It is a collection of different software tools supporting every step of the survey lifecycle, in particular for large international, i.e., multicultural, multinational and multilingual survey studies. The individual tools cover questionnaire design and implementation, managing the translation workflow (also the TRAPD method), survey coding, fieldwork management and monitoring, interviewer tools, sample management, data collection and data dissemination. Having all these steps and functionalities linked in a suite of different tools facilitates the handling because all steps are interlinked, different files can move from one step to another, and so, e.g., conversion or versioning problems disappear. In the questionnaire design stage, such tools allow for, for example, seamless transmission of the questionnaire from the design team to both the systems programming team and the translation team. In this example, the translated questionnaires can then also be directly imported to the CAWI system by the programmer and thus speeds up fieldwork preparations (DataCTRL survey tool suite - Centerdata EN).

4.9.4. Computer-assisted video interviewing (CAVI)

122. Live video (LV) communication tools, like Zoom, offer many benefits of in-person interviews while reducing costs, as travel is unnecessary. This mode of interviewing can be applied not only to data collection, but to in-depth and cognitive interviews and even pretests during the questionnaire design process. Evidence indicates that LV interviewing can yield high-quality data without face-to-face interaction, and recent research notes the absence in LV interviewing of significance interviewer effects, which is a form of error that results when the respondent is, often unconsciously, influenced by the interviewer when responding to survey questions (West et al., 2022). Access to the necessary technology, reliable electricity, and a stable Internet connection are lie to the Internet and stable connectivity are likely to be the most critical barriers to widespread adaptation of LV for questionnaire evaluation and testing, but among some target populations, LV offers a low-cost alternative.

4.9.5. Cognitive interviewing through web-probing

123. Web-probing is a version of a pretesting approach called "embedded probing", where cognitive probes are included in the questionnaire itself rather than in a stand-alone cognitive interview. Increased Internet availability and the relatively low cost of conducting web surveys has led to a resurgence in incorporation of web probes into questionnaires, generally during the pretesting stage of questionnaire development. Evidence indicates that while web probes do not generate as much data as traditional cognitive interviews, respondents do provide open-ended responses to the probes, resulting in data are comparable and that can be used to further assess survey questions (Fowler and Willis, 2019; see also Scanlon, 2019). While application of web probing as an alternative questionnaire testing method has limitations similar to those noted in Section 4.9.4 for LV, this approach has significant promise, depending on the target population.

4.9.6. Use of crowdsourcing for questionnaire design

124. Crowdsourcing in questionnaire design involves leveraging the collective intelligence and diverse perspectives of a large group of people, typically via the internet, to create, refine, and validate survey questions. Social media sites and platforms like Amazon Mechanical Turk (MTurk) can be particularly useful in this process, as they provide access to a diverse pool of respondents which can help to identify cultural, regional, or demographic nuances that might affect how questions are interpreted. Additionally, crowdsourcing can be more efficient and cost-effective compared to traditional methods of survey design, as it allows for rapid collection of feedback from a large number of participants.

125. In a discussion of emerging pretesting approaches, Willis (2019) notes the opportunities provided by social media for "one possibility is that social media content could be seen as constituting… a very large naturalistic, 'big-data focus group'", from which data could be mined, in lieu of typical focus groups (p. 18). Early research on the efficacy of cognitive pretesting compared to crowdsourcing methods, including MTurk and the social media site Facebook suggest that crowdsourcing may be a viable option to test straightforward survey questions which do not require more complex probing by a cognitive interviewer (Edgar & Keating, 2016).

4.10. Resources

126. Table 4.4 lists examples of data collection software to illustrate the variety of tools available for self-administered and interviewer-administered surveys. Their inclusion is not an endorsement or recommendation, but rather a representation of the type of software that can be used in survey research.[17]

Table 4.4. Examples of available software for electronic data capture.

Software name	Developer	Website	Free / Paid
Blaise	Statistics Netherlands	https://blaise.com/	Paid
CSPro	U.S. Census Bureau	https://www.census.gov/data/software/cspro.html	Free
Kobo Toolbox	Kobo	https://www.kobotoolbox.org/	Free
Open Data Kit (ODK)	Get ODK	https://getodk.org/	Free
SurveyCTO	Dobility	https://www.surveycto.com/	Paid
Survey Solutions	World Bank	https://mysurvey.solutions/en/	Free
SurveyToGo	Dooblo	https://www.dooblo.net/	Paid
REDCap	Vanderbilt University	https://www.project-redcap.org/	Free for nonprofits
Tangerine	RTI International	https://www.tangerinecentral.org/	Paid

[17] See also *Comparative Assessment of Software Programs for the Development of Computer-Assisted Personal Interview (CAPI) Applications (English).* LSMS Guidebook Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/719811587030081431/Comparative-Assessment-of-Software-Programs-for-the-Development-of-Computer-Assisted-Personal-Interview-CAPI-Applications.

4.11. References

Acquadro, C., Patrick, D. L., Eremenco, S., Martin, M. L., Kuliś, D., Correia, H., & Conway, K. (2018). Emerging good practices for translatability assessment (TA) of patient-reported outcome (PRO) measures. Journal of patient-reported outcomes, 2, 1-11.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, *20*(1), 139-181.

Axinn, W. G., Pearce, L. D., & Ghimire, D. (1999). Innovations in life history calendar applications. *Social Science Research*, *28*(3), 243-264.

Behr, Dorothée, and Michael Braun. 2023. "How does back translation fare against team translation? An experimental case study in the language combination English-German." Journal of Survey Statistics and Methodology 11 (2): 285-315. doi: https://doi.org/10.1093/jssam/smac005.

Beatty, P. C., Collins, D., Kaye, L., Padilla, J. L., Willis, G. B., & Wilmot, A. (Eds.). (2019). Advances in questionnaire design, development, evaluation and testing. John Wiley & Sons.

Beatty, P.C. & Willis, G.B. (2007). Research Synthesis: The Practice of Cognitive Interviewing. *Public Opinion Quarterly, 71*, 287– 311.

Behr, D., and Braun, M. 2023. "How does back translation fare against team translation? An experimental case study in the language combination English-German." Journal of Survey Statistics and Methodology 11 (2): 285-315. doi: https://doi.org/10.1093/jssam/smac005.

Behr, Dorothée, Dept, S., and Krajĉeva, E. 2019. "Documenting the Survey Translation and Monitoring Process." In Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts (3MC), edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke Stoop, and Brita Dorer, 341-356. New York: John Wiley & Sons.

Behr, D., and Shishido, K.. 2016. "The translation of measurement instruments for cross-cultural surveys." In *The SAGE Handbook of Survey Methodology*, edited by Christof Wolf, Dominique Joye, Tom W. Smith, and Yang-chi Fu, 269-287. London: Sage.

Behr, D., & Zabal, A. (2020). Documenting Survey Translation. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines). DOI: 10.15465/gesis-sg_en_035

Behr, D. 2017. "Assessing the use of back translation: the shortcomings of back translation as a quality testing method." International Journal of Social Research Methodology 20 (6): 573-584. doi: https://doi.org/10.1080/13645579.2016.1252188.

Bernardi, R. A. (2006). Associations between Hofstede's cultural constructs and social desirability response bias. *Journal of Business Ethics*, *65*(1), 43–53.

Beullens, K., Loosveldt, G., & Vandenplas, C. (2017). Classes of nonrespondents in the ESS: which classes are prioritized and which classes should be prioritized in order to reduce nonresponse bias. ESS ERIC work programme.

Blom, A., Doerr, L. A., Cheung, G. Q., Liu, Y., & Harkness, J. (2006). Translations and interpreting: what do survey organizations do?. Survey Res Newslett Survey Res Lab, 37, 1-6.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin*, *119*(1), 111.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design—For market research, political polls, and social and health questionnaires*. San Francisco, CA: Jossey-Bass.

Buskirk, T. D., & Kirchner, A. (2020). Why machines matter for survey and social science researchers: Exploring applications of machine learning methods for design, data collection, and analysis. *Big data meets survey science: A collection of innovative methods*, 9-62.

Chauchard, S. (2013). Using MP3 players in surveys: The impact of a low-tech self-administration mode on reporting of sensitive attitudes. *Public Opinion Quarterly*, *77*(S1), 220–231.

Conrad, F. G., & Schober, M. F. (2021). Clarifying question meaning in standardized interviews can improve data quality even though wording may change: a review of the evidence. *International Journal of Social Research Methodology*, *24*(2), 203-226.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Newbury Park, CA: Sage Publications.

Couper, M. P. (2008). Technology and the survey interview/questionnaire. *Envisioning the survey interview of the future*, 58-76.

Couper, M. P., Beatty, P., Hansen, S. E., Lamias, M., & Marvin, T. (2000). *CAPI design recommendations*. Report submitted to the Bureau of Labor Statistics.

Davis, W. & DeMaio, T. (1993). *Comparing the think-aloud interviewing technique with standard interviewing in the redesign of a dietary recall questionnaire*. Alexandria, VA: ASA Section on Survey Research Methods.

Dean, E., Caspar, R., McAvinchey, G., Reed, L., & Quiroz, R. (2007). Developing a low-cost technique for parallel cross-cultural instrument development: The question appraisal system (QAS-04). *International Journal of Social Research Methodology*, *10*(3), 227-241.

de Jong, J. (2016). Data Collection. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved December 8, 2024, from http://ccsg.isr.umich.edu/.

DeMaio, T., & Landreth, A. (2003, October 21-23). Examining Expert Reviews as a Pretest Method. In P. Prufer, M. Rexroth & F. J. Fowler Jr. (Eds.), QUEST 2003: Proceedings of the 4th Conference on Questionnaire Evaluation Standards (pp. 60-66). Nachrichten: ZUMA

DeMaio, T. & Landreth, A. (2004). Do Different Cognitive Interview Techniques Produce Different Results? In S. Presser et al. (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 891–08). Hoboken, NJ: John Wiley and Sons.

Dillman, D. A. (2019). Asking the right questions in the right way: Six needed changes in questionnaire evaluation and testing methods. *Advances in questionnaire design, development, evaluation and testing*, 25-45.

Dillman, D. A., Gertseva, A., & Mahon-Haft, T. (2005). Achieving usability in establishment surveys through the application of visual design principles. Journal of Official Statistics, 21, 183-214.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). Internet, mail, and mixed-mode surveys: The tailored design method. (3rd ed.). New York, NY: John Wiley & Sons.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method. John Wiley & Sons.

Dorer, Brita. 2020. Advance translation as a means of improving source questionnaire translatability? Findings from a think-aloud study for French and German. Berlin: Frank & Timme.

Dorer, Brita. 2023. "Advance translation: The remedy to improve translatability of source questionnaires? Results of a think-aloud study." Field Methods 35 (1): 33–47. doi: https://doi.org/10.1177/1525822X211072343.

Edgar, J., Murphy, J., & Keating, M. (2016). Comparing traditional and crowdsourcing methods for pretesting survey questions. *Sage Open*, *6*(4), 2158244016671770.

European Social Survey (2022). ESS Round 11 Translation Guidelines. London: ESS ERIC Headquarters. Information for the Second National Co-ordinators Meeting

European Social Survey (2024). ESS Round 12 Survey Specification for ESS ERIC Member, Observer and Guest Countries (ESS012_projection_specification_v2.pdf (europeansocialsurvey.org))

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. Journal of Official Statistics, 27(4), 569-599.

Fitzgerald, R., & Zavala-Rojas, D. (2020). A Model for Cross-National Questionnaire Design and Pretesting. Advances in questionnaire design, development, evaluation and testing, 493-520.

Forsyth, B., Rothgeb, J., & Willis, G. (2004). Does Questionnaire Pretesting Make a Difference? An Empirical Test Using a Field Survey Experiment. In S. Presser, et al. (Eds.), *Questionnaire Development, Evaluation, and Testing* (pp. 525–546). Hoboken, NJ: John Wiley and Sons.

Fowler, F. J. Jr. (1995). *Improving survey questions: Design and evaluation* (Vol. 38, Applied social research methods series). Thousand Oaks, CA: Sage Publications.

Fowler, F. J. (2014). Survey Research Methods (5th ed.). SAGE Publications.

Fowler, S., & B. Willis, G. (2019). The practice of cognitive interviewing through web probing. *Advances in questionnaire design, development, evaluation and testing*, 451-469.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological methodology*, 37-68.

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. Review of general psychology, 15(4), 380-387.

Geisen, E., & Murphy, J. (2019). A compendium of web and mobile survey pretesting methods. *Advances in questionnaire design, development, evaluation and testing*, 287-314.

Gerber, E. R. (1999). The view from anthropology: Ethnography and the cognitive interview. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, and R. Tourangeau (Eds.). Cognition and survey research (pp. 217-234). New York, NY: Wiley.

Gibson and Louw, ND. https://www.povertyactionlab.org/resource/survey-programming?lang=fr

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID) a web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3-22.

Groves, Robert M., et al. *Survey Methodology*. Second edition., Wiley, 2009.

Hadler P., Neuert C.E., Ortmanns V., & Stiegler A. (2022). "Are You…? Asking Questions on Sex with a Third Category in Germany." Field Methods, Vol. 34(2) 91–107, DOI: 10.1177/1525822X211072326

Hansen, S. E. & Couper, M. P. (2004). Usability testing as a means of evaluating computer assisted survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley & Sons.

Hansen, S. Lee, H.J., Lin, Y., & McMillan, A. (2016). Instrument Technical Design. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved December 8, 2024, from http://ccsg.isr.umich.edu/.

Harkness, J. A., & Schoua-Glusberg, A. (1998). Questionnaires in Translation. In J. A. Harkness (Ed.), Cross-Cultural Survey Equivalence (Vol. 3). Mannheim: Zentrum für Umfragen, Methoden und Analysen-ZUMA.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. Van de Vijver, & P. P. Mohler (Eds.), Cross-cultural survey methods (Vol. 325, pp. 35). Wiley-Interscience Hoboken, NJ.

Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Questionnaire Translation and Assessment. In S. Presser, J. Rothgeb, M. Couper, J. Martin, & E. Singer (Eds.), Methods for Testing and Evaluating Survey Questionnaires. New Jersey: John Wiley and Sons.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010a). Designing questionnaires for multipopulation research. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell, & T. Smith (Eds.). Survey methods in multinational, multiregional, and multicultural contexts (pp. 31-57). Wiley-Blackwell.

Harkness, J. A., Villar, A., & Edwards, B. (2010b). Translation, Adaptation, and Design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), Survey methods in multinational, multiregional, and multicultural contexts (pp. 117-140). Wiley-Blackwell.

Hauck, W. 2004. Ubersetzer in ¨ offentlichen Verwaltungen als Mitdenker des Autors ¨ — Viel ungenütztes Potenzial. In Internationales CIUTI-Forum, marktorientierte translationsausbildung, eds. M. Forstner and H. Lee-Jahnke, 211–24. Bern: Peter Lang

Hewett, P. C., Mensch, B. S., & Erulkar, A. S. (2004). Consistency in the reporting of sexual behaviour by adolescent girls in Kenya: A comparison of interviewing methods. *Sexually Transmitted Infections*, 80(Suppl 2), 43–48.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. Public opinion quarterly, 67(1), 79-125.

ICF. (2024). *Feed the Future Survey Implementation Document: Translation Protocol.* Reston, VA: ICF.

Ikart, E. M. (2019). Survey questionnaire survey pretesting method: An evaluation of survey questionnaire via expert reviews technique. Asian Journal of Social Science Studies, 4(2), 1.

Jansen, H. & Hak, T. (2005). The Productivity of the Three-Step Test-Interview (TSTI) Compared to an Expert Review of a Self-Administered Questionnaire on Alcohol Consumption. *Journal of Official Statistics*, *21*, 103–120.

Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., & Park, D. S. (2023). Exhaustive or exhausting? Evidence on respondent fatigue in long surveys. *Journal of Development Economics*, *161*.

Ji, L. J., Schwarz, N., & Nisbett, R. E. (2000). Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and social psychology bulletin*, *26*(5), 585-593.

Knauper, B., Schwarz, N., Park, D., & Fritsch, A. (2007). The perils of interpreting age differences in attitude reports: Question order effects decrease with age. Journal of Official Statistics, 23(4), 515.

Kreuter, F., Yan, T., & Tourangeau, R. (2008). Examining the Effectiveness of Latent Class Analysis to Item Evaluation. *Journal of the Royal Statistical Society*, *171*(Series A), 723–738.

Krosnick and Presser, 2010. Emerald_HSR-V017_9 263..313 (stanford.edu)

Krosnick, J. A. (2018). Improving question design to maximize reliability and validity. The Palgrave handbook of survey research, 95-101.

Lalwani, A. K., Shavitt, S., & Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *Journal of Personality and Social Psychology*, *90*(1), 165–178.

Lara, D., Strickler, J., Olavarrieta, C. D., & Ellertson, C. (2004). Measuring induced abortion in Mexico A comparison of four methodologies. *Sociological Methods & Research*, *32*(4), 529-558.

Lipps, O., & Ochsner, M. (2018). Additional languages and representativeness. Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC), 859-877.

Mangione, T., Fowler, F. J., & Oksenberg, L. (1992). Question characteristics and interviewer effects. *Journal of Official Statistics*, *8*(3), 293–307.

Mensch, B. S., Hewett, P. C., & Erulkar, A. S. (2003). The reporting of sensitive behavior by adolescents: a methodological experiment in Kenya. *Demography*, *40*(2), 247–268.

Metheney, E. A., & Yehle, L. (2024). Exploring the Potential Role of Generative AI in the TRAPD Procedure for Survey Translation. arXiv preprint arXiv:2411.14472.

Miller JD. (1984). *A new survey technique for studying deviant behavior*. Unpublished Doctoral Dissertation, George Washington University.

Miller, K. (2019). Conducting cognitive interviewing studies to examine survey question comparability. In Johnson, T. P., Pennnell, B-E., Stoop, I. A. L., & Dorer, B. (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 203–225). Hoboken, NJ: John Wiley & Sons.

Morgan, D. L. (1996). Focus groups. *Annual review of sociology*, *22*(1), 129-152.

Nichols, E., Olmsted-Hawala, E., Holland, T., & Riemer, A. A. (2019). Usability Testing Online Questionnaires: Experiences at the US C ensus B ureau. *Advances in Questionnaire Design, Development, Evaluation and Testing*, 315-348.

Olivos, F., & Liu, M. (2024). ChatGPTest: opportunities and cautionary tales of utilizing AI for questionnaire pretesting. *Field Methods*.

Ornstein, Michael D. A Companion to Survey Research. SAGE, 2013.

Peytcheva, E. (2020). The Effect of Language of Survey Administration on the Response Formation Processes. In M. Sha & T. Gabel (Eds.), *The essential role of language in survey research*.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. Methods for testing and evaluating survey questionnaires, 1-22.

Prewitt-Freilino, J. L., Caswell, T. A., & Laakso, E. K. (2012). The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. Sex roles, 66(3), 268-281.

Rasinski, K. A., Lee, L., & Krishnamurty, P. (2012). Question order effects. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics (pp. 229–248). American Psychological Association.

Reeve, B. & Mâsse, L. (2004). Item Response Theory (IRT) modeling for questionnaire evaluation. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley and Sons.

Repke, L, and Dorer, B. 2021. "Translate wisely! An evaluation of close and adaptive translation procedures in an experiment involving questionnaire translation." International Journal of Sociology 51 (2): 135-162. doi: https://doi.org/10.1080/00207659.2020.1856541.

Rossi, P. & Anderson, A. (1982). The factorial survey approach: An introduction. In P. Rossi and S. Nock (Eds.), *Measuring social judgments: The factorial survey approach*. Thousand Oaks, CA: Sage Publications.

Rothgeb, J., Willis, G., & Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results? *Proceedings of the Section on Survey Methods*. Alexandria, VA: American Statistical Association.

Saris, W., van der Veld, W., & Gallhofer. I. (2004). Development and improvement of questionnaires using predictions of reliability and validity. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley and Sons.

Scanlon, P. (2019). Using targeted embedded probes to quantify cognitive interviewing findings. *Advances in questionnaire design, development, evaluation and testing*, 427-449.

Schaad et al. (2020). Improving the Question Appraisal System (QAS): Moving Further Away from Black Magic and Black Boxes. http://www.asasrms.org/Proceedings/y2020/files/1505493.pdf.

Schoua-Glusberg, A., & Villar, A. (2014). Assessing translated questions via cognitive interviewing. Cognitive interviewing methodology, 51-67.

Schaeffer, N.C. and Dykema, J. (2011). Response 1 to Fowler's chapter: coding the behavior of interviewers and respondents to evaluate survey questions. In: Question Evaluation Methods: Contributing to the Science of Data Quality (eds. J. Madans, K.A. Miller and G.Willis), 23–39. Hoboken, NJ:Wiley.

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In J.A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 175–190). John Wiley & Sons, Inc.

Schwarz, N., Hippler, H. J., & Noelle-Neumann, E. (1992). A cognitive model of response-order effects in survey measurement. In Context effects in social and psychological research (pp. 187-201). New York, NY: Springer New York.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. American psychologist, 54(2), 93.

Smith, L. C., Dupriez, O., & Troubat, N. (2014). Assessment of the reliability and relevance of the food data collected in national household consumption and expenditure surveys. *International Household Survey Network*, *82*.

Sudman, S. & Bradburn, N. M. (1982), *Asking questions: A practical guide to questionnaire design.* San Francisco, CA: Jossey-Bass.

Tarnai, J. & Moore, D. (2004). Methods for testing and evaluating computer-assisted questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires*. New York, NY: John Wiley and Sons.

Toninelli, D., & Revilla, M. (2019). How mobile device screen size affects data collected in web surveys. *Advances in questionnaire design, development, evaluation and testing*, 349-373.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological bulletin*, *103*(3), 299.

Tourangeau R., Rips L. J., Rasinski K. (2000), The Psychology of Survey Responses, New York: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, *133*(5), 859.

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, *26*(2), 169-181.

Tourangeau, R., Maitland, A., Steiger, D., and Yan, T. (2019). A framework for making decision about question evaluation methods. *Advances in questionnaire design, development, evaluation and testing*, 47-73.

Triandis, H. C. (1995). *Individualism and collectivism*. Westview Press.

Uskul, A. K., & Oyserman, D. (2006). Question Comprehension and Response: Implications of Individualism and Collectivism. In Y.-R. Chen (Ed.), *National Culture and Groups* (Vol. 9, pp. 173–201). Emerald Group Publishing Limited.

Uskul, A. K., Oyserman, D., & Schwarz, N. (2010). *Cultural emphasis on honor, modesty, or self-enhancement: Implications for the survey response process*.

Waickman, C. R., & Corbett, A. (2022) Multilingual Communities Require Multilingual Surveys: A Language Justice-Informed Approach to the New York City Housing and Vacancy Survey. Paper presented at the 77th conference of the American Association for Public Opinion Research (AAPOR).

West, B. T., Ong, A. R., Conrad, F. G., Schober, M. F., Larsen, K. M., & Hupp, A. L. (2022). Interviewer effects in live video and prerecorded video interviewing. *Journal of Survey Statistics and Methodology*, *10*(2), 317-336.

Willimack, D.K., Lyberg, L. E., Martin, J., Japec, L., & Whitridge, P. (2004). Evolution and adaptation of questionnaire development, evaluation, and testing methods for establishment surveys. In S. Presser, J. M. Rothger, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 385–408). Hoboken, NJ: John Wiley & Sons, Inc.

Willis, G. B. (2005). *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Willis, G.B. (2019). Questionnaire design, development, evaluation, and testing: Where are we, and where are we headed? *Advances in questionnaire design, development, evaluation and testing*, 1-23. Beatty, P., Collins, D., Kaye, L., Padilla, J.L., Willis, G., & Wilmot, A. Wiley.

Willis, G. B., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology*, *5*(3), 251-267.

Willis, G. B. & Lessler, J. T. (1999). *Question appraisal system BRFSS-QAS: A guide for systematically evaluating survey question wording* (Report prepared for CDC/NCCDPHP/Division of Adult and Community Health Behavioral Surveillance Branch). Rockville, MD: Research Triangle Institute. Retrieved April 14, 2010, from http://appliedresearch.cancer.gov/areas/cognitive/qas99.pdf

Wilson, L., & Dickinson, E. (2021). Respondent centred surveys: Stop, listen and then design. Sage Publishing.

World Bank. (2024). Survey Pilot - Dimewiki

Yan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating Survey Questions: A Comparison of Methods. *Journal of Official Statistics*, *28*(4), 503–529.

Yang, Y., Harkness, J. A., Chin, T., & Villar, A. (2010). Response styles and culture. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 203–223). John Wiley & Sons.

Zavala-Rojas, Diana, Dorothée Behr, Brita Dorer, Danielly Sorato, and Veronika Keck. 2024. "Using Machine Translation and Post-editing in the TRAPD approach: effects on the quality of translated survey texts." Public Opinion Quarterly 88 (1): 123-148. https://doi.org/10.1093/poq/nfad060.

CHAPTER 5
FRAMES AND SAMPLING

Juan Pablo Ferreira (National Institute of Statistics, Uruguay)
Piotr Jabkowski (Adam Mickiewicz University)
Dale A. Rhoda (Biostat Global Consulting)

CHAPTER 5
FRAMES AND SAMPLING

5.1. Introduction

1. This chapter explores key concepts related to sampling frames and sample selection methods, particularly in surveys targeting households and individuals. A sampling frame refers to a well-defined list or database of units from the target population accessible for sample selection. In contrast, a sampling design refers to the strategy or method used to select those units from the sampling frame. Both elements are essential for ensuring that the sample drawn for the survey is representative of the target population, which is crucial for obtaining accurate and reliable estimates. By employing probability sampling techniques, National Statistical Offices (NSOs) can create samples that allow valid statistical inferences while minimizing errors and biases in the results.

2. Effective sampling ensures that the data collected from sampled units can accurately represent the entire population without conducting a census, which is critical when the resources required for a census may be prohibitively costly and time-intensive.

3. At the core of sampling methodology lie several fundamental goals that guide the design and execution of surveys:

- Unbiased Estimates: The expected value of the survey estimate should equal the unknown population parameter, which ensures that survey results are systematically accurate rather than skewed high or low.
- Precision: Precision refers to the sampling variability or expected deviation of the estimate from the population parameter. Survey estimates must fall within an acceptable range for decision-making or action.
- Timeliness: The speed of data collection and analysis must align with stakeholders' operational and decision-making needs.
- Affordability: Survey costs must remain within budgetary constraints while still delivering results of acceptable quality.

4. Because surveys must typically meet unbiased and precise objectives, most surveys rely on probability sampling. Probability sampling means that:

[1] Every member of the population under study has a non-zero probability of being selected into the survey sample.

[2] It is possible to calculate the probabilities for those who are selected.

5. A foundational principle of survey research is that conclusions drawn from probability samples are:

- Representative of the broader population: Probability sampling ensures that inferences made from the sample can be generalized to the entire population.
- Unbiased (or nearly so): Survey estimators are designed to avoid systematic deviations from the true value of the population parameter.

- Precise: Sampling variability, often referred to as sampling error, is quantifiable and bounded within probabilistic limits.

Understanding these principles is fundamental to designing and interpreting sample surveys, forming the basis for reliable, actionable insights.

6. Throughout the chapter, strategies for maintaining and updating sampling frames are discussed, along with how these decisions impact the quality of survey estimates. Over-coverage (inclusion of ineligible units) and under-coverage (exclusion of eligible units) are two common issues that can bias survey results. Both problems highlight the critical role of accurate and up-to-date sampling frames in maintaining data integrity and representativeness.

7. Drawing from practical examples and lessons learned from NSOs and fieldwork agencies, this chapter provides a detailed guide for selecting and managing sampling frames, emphasizing the importance of adapting them to the dynamic realities of populations. It also explores new sources for creating sampling frames, such as administrative records, telephone-based frames, and other data sources, while addressing coverage and quality control challenges.

8. Additionally, various types of sampling frames used in practice are presented, along with their advantages, disadvantages, and real-world applications. Sampling frames provide the foundation for sample selection, but the overall quality of survey estimates depends on multiple factors, including the robustness of the sample design, the accuracy of the sampling frame, and the operational execution of the survey.

9. The chapter also outlines key sample designs commonly employed in survey research, including:

- Simple random sampling
- Stratified sampling
- Systematic sampling
- Cluster sampling
- Multi-stage sampling
- Multi-phase sampling

Each design is discussed regarding its applications, strengths, and limitations, particularly how they are adapted to diverse population structures and survey objectives.

10. A critical component of survey design is selecting the sample size. This chapter explores the theoretical and practical considerations of sample size determination, including balancing statistical precision, cost constraints, and operational feasibility. It also addresses the importance of incorporating design effects and accounting for complex sampling structures when calculating sample sizes.

11. In conclusion, this chapter guides the selection of a practical combination of sampling frame and design, ensuring that survey objectives are met affordably with accuracy, and reliability. By understanding the principles of sampling and applying robust methodologies, researchers can collect high-quality data to support evidence-based decision-making.

5.2. Sampling frames

5.2.1. Key concepts in sampling frames

12. The following paragraphs focus specifically on *sampling frames*, a fundamental component of survey design. A *sampling frame* refers to a well-defined set of units from the target population accessible for sample selection (Särndal, Swensson, and Wretman 1992). Within the broader context of survey design,

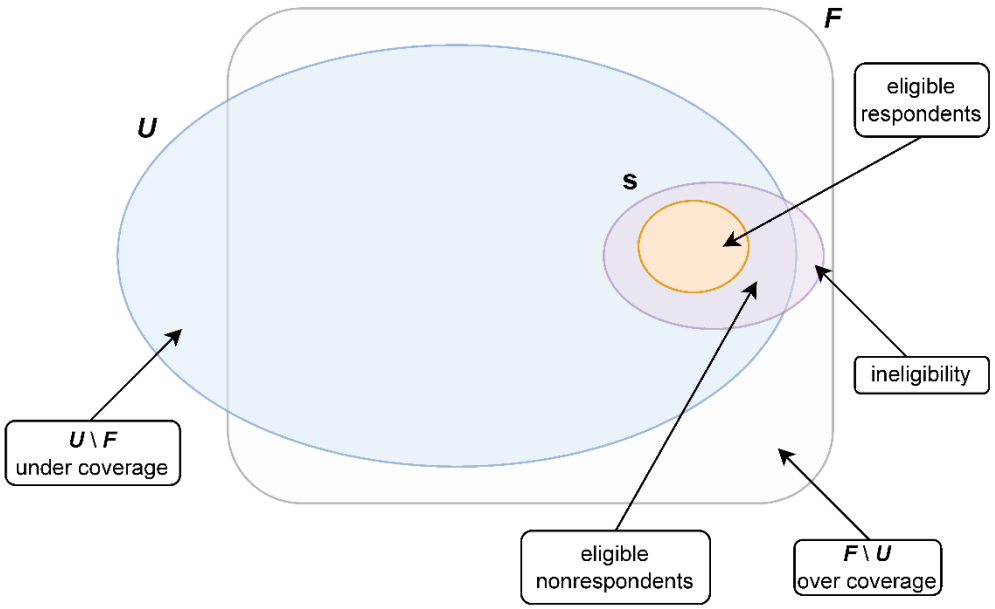
the *sampling frame* serves as the operational list from which the sample is drawn, while the *sample design* defines the method or strategy used to select units from this frame. Both elements play distinct yet interconnected roles in ensuring the quality and representativeness of survey estimates. Errors in the sampling frame can result in systematic non-sampling errors, such as under-coverage of crucial population subgroups (Biemer et al. 2017; S. L. Lohr 2008). Under-coverage may occur if specific units from the population are not part of the sampling frame but possess distinct characteristics. For example, if a survey aims to estimate the *average household income* (the parameter of interest) and new informal housing units are not included in the sampling frame, this omission can result in overestimation of the average income.

13. This section outlines best practices for constructing and using sampling frames across different stages of the survey cycle, emphasizing strategies for improving the sample selection process, increasing response and eligibility rates, and constructing sample weights for the estimation stage.

14. In surveys of individuals and households, the target population can include households, individuals, or sometimes dwellings, depending on the survey's objectives. For example, surveys focused on household income, expenditure, or facilities may define the household as the target population unit. In contrast, other surveys may focus on individuals living within households as the primary units of analysis. Clearly defining the target population is essential for aligning survey objectives with sampling strategies.

15. Differences or mismatches between these sets can impact the accuracy and validity of survey estimates. **Figure 5.1** outlines differences between various concepts related to the frames for sampling, and the interplay between those concepts impacts the overall research process and the validity of survey results.

Figure 5.1. Concept of population (**U**), sampling frame (**F**), and sample (**s**).



16. The *target population* consists of all elements (e.g., households or individuals) that make up the group of interest. The *sampling frame* represents the operational list from which the sample is drawn. The *sampled units* may include specific individuals, households, dwellings, or other entities selected—usually

randomly under complex sampling designs—to participate in the survey. The *eligible respondents* are those selected units that fall within the scope of the survey and ultimately participate by providing valid information on the variables of interest. It is important to note that some elements selected for the sample may be ineligible, meaning they do not belong to the target population but are mistakenly included in the frame due to quality issues, particularly *over-coverage*.

17. **Key Concepts** include:

- **Target population (*U*):** All elements (e.g., households or individuals) that constitute the Universe of interest.
- **Sampling frame (*F*):** The operational list from which the sample is drawn. Its quality directly affects survey accuracy.
- **Sampled units (*s*):** Specific units selected from the frame - individuals, households, or dwellings - to participate in the survey.
- **Eligible respondents (*ER*):** Selected units who turn out to fall within the survey's scope and who provide valid data for analysis.
- **Eligible nonrespondents (*ENR*):** Selected units who fall within the scope of the survey but do not provide valid data. This nonresponse may occur for various reasons, including refusal to participate, inability to establish contact, or failure to complete the interview.

5.2.1.1. Target population

18. The term *target population* is used in survey research to define the individuals, households, or other units a study aims to represent. In other words, the target population consists of the units from which conclusions are intended to be drawn and to which survey results are meant to be generalized. Clearly defining the target population is a critical step in the survey process, as it establishes the boundaries of the research and determines the eligibility of potential participants. A well-defined target population includes explicitly outlined inclusion criteria, such as geographic location, temporal parameters, and specific demographic and social characteristics relevant to the study's objectives.

19. For example, a survey targeting the general population might define its target population as all adults within a specific country while allowing for certain exclusions, such as individuals residing in institutions like prisons, hospitals, military bases, or sparsely populated areas that are difficult and costly to access. These exclusions are often made for practical reasons but must be explicitly defined to ensure clarity and accuracy in the research process. Precision in defining the target population is essential to avoid specification errors, enabling survey findings to be appropriately generalized to the population of interest and minimizing the potential for bias or misinterpretation.

5.2.1.2. Sampling frame and frame units

20. A *sampling frame* can be defined as a list or database from which a sample of units belonging to the target population is drawn to conduct a survey. It represents the physical or operational source from which the sample is selected and is a practical means of accessing the target population. While the sampling frame is intended to represent the target population, it does not always consist of the same type of units. For example, the target population might consist of individuals, but the sampling frame could be defined as geographic areas, dwellings, or households that serve as proxies or access points to reach those individuals.

21. A sampling frame must guarantee that each member of the population has a non-zero probability of selection:

$$\pi_k = P[k \in sample] > 0 \; \forall \; k \in U \tag{5-1}$$

This requirement ensures the representativeness of the entire population and allows for unbiased estimates to be obtained using, for example, the *Horvitz-Thompson estimator* (Horvitz and Thompson 1952; Bethlehem 2002) (*see Chapter 8*).

22. The term *frame units* or *sampling units* refers to the elements listed within the sampling frame. A frame may include various types of sampling units, depending on the survey design. These units could be individuals, households, blocks, addresses, geographic areas, census units, or other relevant entities that provide a structured means to access or represent the target population.

5.2.1.3. Sampled units

23. *Sampled units* are the specific elements selected from a *sampling frame* to participate in a survey. These units constitute a subset of the target population and are chosen to gather data for generalization to the entire population. As the *sampling frame* serves as the operational list from which these units are drawn, the sampling statistician must ensure that each unit has a non-zero probability of being selected. This property is essential for guaranteeing that the resulting sample is randomly drawn and representative of the population. Additionally, the sample must be selected using a rigorously random probability sampling design at each stage of selection to maintain statistical validity and limit bias.

5.2.1.4. Surveyed units

24. The term *surveyed units* or *eligible respondents (ER)* refers to a subset of selected units that were successfully contacted and participated in the survey, providing valid responses for analytical purposes. In contrast, *nonrespondents* are sampled units that did not participate in the survey. *Unit nonresponse* can occur for several reasons, including *non-contact* (i.e., inability to reach the selected unit after repeated attempts), *refusals* (when contacted individuals decline to participate), or other factors such as illness, language barriers, or inability to respond (see Chapter 6).

25. A further distinction exists between two categories of nonrespondents: *eligible nonrespondents* and *ineligible units*. *Eligible nonrespondents* meet the inclusion criteria for the study but fail to provide data (AAPOR 2023). Determining eligibility among nonrespondents can often be challenging due to limited or missing information.

26. On the other hand, *ineligible units* do not meet the inclusion criteria for the survey but may still be included in the sampling frame—their inclusion, whether due to frame imperfections or other reasons, can result in *over-coverage errors*.

27. The distinction between respondents and nonrespondents is crucial for calculating survey **outcome rates**, including **response rates**, **contact rates**, and **refusal rates** (*for details, consult AAPOR, 2023, or see Chapter 6*). These rates are essential for assessing the potential for *nonresponse bias* and determining the extent to which the surveyed units represent the entire target population.

28. To address nonresponse, researchers may implement strategies such as weighting adjustments, follow-up surveys, or incentives to encourage participation and mitigate the effects of missing data.

5.2.2. Role of the sampling frame in the GSBPM context

29. The sampling frame plays a crucial role in statistics derived from sample surveys and aligns with key stages of the Generic Statistical Business Process Model (GSBPM) (United Nations Economic Commission for Europe 2019). This section examines how the sampling frame connects with specific phases and subprocesses of the GSBPM, providing a structured perspective on its integration into the statistical lifecycle.

30. In the Build phase, subprocess 3.2 (Build and Maintain Frames and Registers) focuses on establishing and managing the sampling frame and statistical registers. This subprocess involves creating a reliable and comprehensive list of units from the target population, which serves as the foundation for sample selection. Additionally, subprocess 3.3 (Design the Sample) emphasizes the importance of using the sampling frame effectively during the sampling design stage, ensuring alignment with survey objectives and population characteristics.

31. In the Process phase, subprocess 7.2 (Update Statistical Registers and Frames) highlights the necessity of periodically reviewing and updating the sampling frame. These updates address changes in the target population, structural shifts, or emerging inaccuracies, ensuring that the sampling frame remains relevant and reliable over time.

32. A well-maintained and regularly updated sampling frame supports accurate coverage of the target population, reduces the risk of non-sampling errors, and strengthens the foundation for robust statistical inference across the entire statistical process. By aligning with these specific subprocesses of the GSBPM, the sampling frame contributes to the overall quality, transparency, and reliability of survey-based statistical outputs.

5.2.3. Types of sampling frames

33. The quality of a sample in survey research depends on the comprehensiveness of the sampling frame. As previously mentioned, a sampling frame is a list that provides access to the elements of the target population, enabling sample selection. The degree of representativeness and reliability of the sample is determined by how thoroughly the sampling frame encompasses the entire target population. Selecting an appropriate sampling frame influences survey quality, constrains variations in selection probabilities, and directly impacts the design effect, which determines the efficiency of the sample and the required sample size for achieving reliable results. Therefore, it is essential to carefully select an appropriate sampling frame at the outset of the research process.

34. Sampling frames can be categorized based on the type of unit included in the frame (e.g., persons, dwellings, telephone numbers, geographic areas) rather than their sources (e.g., census data, administrative records, or combinations of these). Each frame type has distinct characteristics, advantages, and limitations, and the choice depends on the survey's objectives, population coverage, and available data sources.

35. **Person-level frames.** A person-level frame consists of a list of individuals from which a sample can be drawn. These frames are often derived from population registers or recent census data. Population registers typically provide a comprehensive and up-to-date list of residents within a defined geographical area, ensuring high coverage of the target population. Sampling directly from individuals in such registers is generally more efficient, resulting in lower standard errors of estimates compared to multi-stage sampling designs. In multi-stage designs, intermediate sampling units (e.g., blocks, households) are selected before reaching the final target individuals.

36. **Dwelling/household frames or address/building lists.** Dwelling, household, address, and building are distinct terms, and their precise definitions should guide their usage:

- *Dwelling:* A physical space intended for residential use.
- *Household:* A social unit, typically consisting of individuals living together and sharing resources.
- *Address:* A formal descriptor of a dwelling's location.
- *Building:* A physical structure containing one or more dwellings or addresses.

Dwelling or household frames and address or building lists are often treated as interchangeable terms, but distinctions exist. Dwelling or household frames focus on residential units or groups of people sharing resources, while address or building lists emphasize physical locations or structures. In the absence of person-level frames, these frames serve as viable alternatives. Although they require within-household respondent selection, they typically ensure broad population coverage.

37. **Telephone number frames.** Telephone number frames can be constructed using pre-existing lists of phone numbers obtained from administrative sources, census data, or telecommunications providers. While random-digit dialling (RDD) is one method of generating telephone samples, it is not the only approach. Telephone frames allow direct contact with respondents and cover mobile and landline numbers, reaching diverse demographics. However, within-household respondent selection remains challenging when conducted via telephone compared to on-site selection by field staff.

38. **Area sampling with field enumeration.** Area sampling becomes a necessary approach when no pre-existing lists of individuals, addresses, or dwellings are available. In this method, the target area is divided into smaller segments or clusters, often referred to as enumeration areas (EAs). Researchers then conduct fieldwork to enumerate all dwellings or households within these clusters. Despite being labour-intensive and time-consuming, area sampling is frequently the only viable option in regions lacking other types of frames.

39. **List frames from administrative or program data.** In addition to these frames, list frames derived from administrative records or program-specific beneficiary lists can serve as sampling frames. Examples include lists of individuals enrolled in government programs, healthcare registries, or educational databases. While these lists might not offer full population coverage, they provide valuable sampling resources, especially when traditional frames are unavailable.

40. Each sampling frame type has unique strengths and limitations. The choice of frame should consider factors such as coverage, completeness, accuracy, and the specific requirements of the survey. Additionally, the source of the frame (e.g., census, administrative data, program lists) should be explicitly stated to ensure clarity and transparency.

5.2.3.1. Person-level frames/population registers

41. Population registers are widely recognized as one of the most effective sampling frames for survey research due to their accuracy and detailed nature (Stoop et al. 2010). These registers, typically maintained by government agencies, provide information on individuals residing within a country. Their utility stems from two key factors: extensive coverage and the availability of individual-level auxiliary variables, such as age and gender. These variables are valuable for sample design (e.g., stratification with varying sampling rates) and for creating final weights through calibration techniques (e.g., post-stratification). This process helps mitigate biases introduced by nonresponse from selected units (e.g., households or individuals) in the theoretical sample. However, while population registers offer numerous advantages, researchers must consider their limitations and associated challenges.

42. The most significant advantage of population registers is their extensive population coverage. These registers typically include nearly all individuals within a country's borders, providing a comprehensive sampling frame. Such coverage facilitates equal probability sampling, leading to a design effect comparable to simple random sampling, which minimizes the number of sampled units required to achieve an effective sample size.

43. Moreover, population registers often include individual-level auxiliary variables, such as age, gender, geographic region, or other demographic characteristics. These detailed datasets allow for precise stratification, improving the sample's accuracy and representativeness. Stratification based on individual-level characteristics tends to be more effective than relying on broader geographic or regional variables, ultimately enhancing the precision of survey results and reducing sampling errors.

44. Sampling frames derived from population registers generally include contact information (e.g., phone numbers or email addresses), enabling data collection through phone, web, or mail self-completion modes. Consequently, sample designs can often remain straightforward, with the sampling unit aligning directly with the unit of analysis. This approach is typically more efficient than multi-stage sampling designs, which are frequently employed to reduce data collection costs by minimizing geographic dispersion. However, multi-stage sampling usually results in higher design effects (*DEFF*), which indicates how much the sampling design increases the variance of an estimate compared to simple random sampling. A higher *DEFF* means less precise estimates, requiring larger sample sizes to achieve the same level of accuracy.

45. Government entities, such as national statistical agencies, civil registration offices, or local government authorities, commonly maintain population registers. In many countries, these registers are built using administrative records, including birth and death certificates, immigration and emigration data, and other official documentation tracking residency status. Additionally, they may integrate information from national identification systems, voter registries, and social security records to maintain accuracy and currency. Population registers are sometimes linked to other administrative databases, such as healthcare systems, education records, and taxation databases, enriching their detail and reliability. These integrated datasets provide a holistic view of the population and ensure that demographic and socioeconomic information remains up-to-date. However, access to these registers for research purposes varies across countries, with some imposing strict regulations to safeguard individuals' privacy.

46. Despite their numerous benefits, population registers have certain limitations. Over-coverage is a common issue, where individuals no longer residing in the country (e.g., those working or studying abroad for extended periods) remain listed in the register, potentially distorting the sample. Under-coverage is another concern, as certain groups, including recent immigrants or undocumented individuals, may be inadequately represented, leading to samples that fail to capture population diversity entirely and potentially biased results. Furthermore, opt-outs – individuals included in the register who have explicitly requested not to be contacted for research – pose additional challenges. While these individuals should technically remain in the sampling frame and be recorded as refusals if selected, their exclusion can affect the representativeness of the final sample.

47. Population registers are an invaluable resource for survey sampling due to their comprehensive coverage and the depth of auxiliary information they offer. However, researchers must remain vigilant about potential challenges, such as over-coverage, under-coverage, and the impact of opt-outs, to ensure accurate and reliable survey outcomes.

5.2.3.2. Dwelling frames or address lists

48. Dwelling frames or address lists often serve as survey sampling frames, in the common scenario where population registers are unavailable or inaccessible (Jabkowski and Kołczyńska 2020). These frames consist of lists of residential addresses or dwellings/households, enabling researchers to indirectly reach the target population units by selecting dwellings, households, or addresses as primary sampling units (*PSU*). Common sources for such registers include postal delivery services, land registration records, taxation databases, and utility connection records.

49. One of the primary advantages of dwelling frames is their broad population coverage. Since most households are associated with postal addresses or utility services, these lists can provide a comprehensive sampling frame encompassing a large proportion of dwelling units. This accessibility facilitates the implementation of nationwide surveys without requiring overly complex or resource-intensive procedures. Additionally, dwelling frames are often geographically organized, simplifying stratification and ensuring an appropriate geographic distribution of the sample.

50. However, dwelling frames also present important challenges that often require introducing an additional stage of sampling and constructing a new sampling frame at that stage. A significant difficulty lies in selecting equal-probability samples of individuals, which refers to a sampling design where each individual in the target population has an equal chance of being selected. Unlike population registers, which list individuals, dwelling frames list addresses or households with varying numbers of eligible respondents. This variability leads to higher design effects compared to individual-based sampling frames. Specifically, the within-household selection procedure introduces a loss of statistical precision, as the overall selection probabilities of individuals become inversely proportional to the number of eligible residents in each dwelling. This discrepancy often necessitates larger sample sizes to achieve the desired level of precision.

51. To address this issue, it becomes essential to collect information on the number of eligible individuals residing in each household. This data is critical for determining appropriate design weights, adjusting the selection process, and correcting potential biases caused by unequal selection probabilities. Furthermore, in low- and middle-income countries, dwelling frames or address lists are often unavailable, incomplete, or outdated, posing significant challenges for their effective use in survey sampling.

52. In summary, dwelling frames or address lists are valuable tools for surveying target populations when individual-level sampling frames are unavailable. While they offer significant advantages regarding accessibility, geographic coverage, and practicality, they also require careful implementation of within-household selection methods to address unequal selection probabilities (see Section 5.3.12.2, Selecting a respondent within a household). When using dwelling or household frames in surveys, selecting target respondents within the household is critical.

5.2.3.3. Area frames with field enumeration to create a frame of dwellings/households

53. Area frames are widely used in survey sampling, particularly when other sampling frames, such as population registers or address lists, are unavailable or incomplete. While gridded population datasets are one option (Thomson 2022), area frames can also be built from existing geographical or administrative divisions (Hansen and Hauser 1945). This approach involves selecting small geographical areas—such as districts, blocks, or streets—and conducting field enumeration to construct an up-to-date frame of all dwellings or households within the selected PSU areas. Field enumeration is central to ensuring that all households in the selected areas are accounted for, addressing coverage gaps, and enabling accurate sampling.

54. The sources for area frames typically include existing geographical or administrative divisions, which generally cover an entire country's land area. These divisions often have precisely delineated boundaries, are mapped, and have associated population figures. Census data is a common starting point, as census enumeration areas are designed to cover the national territory with well-defined boundaries and mapped locations. These areas often include population figures essential for assigning size measures and calculating selection probabilities. Field enumeration supplements these data by validating or updating household lists in the field, ensuring that each dwelling is accurately recorded.

55. Other sources for area frames may include administrative records, postal service areas, street directories, or custom-defined grids based on maps. Note that the population estimates that accompany area frames are likely to be out-of-date in regions without a recent census, with rapid population changes or with informal settlements. Those regions may be underrepresented in the sample if the selection process uses probability proportional to population. In regions where official records may be outdated or incomplete it may be preferable to construct a new frame based on digital gridded population estimates that use up-to-date data inputs and more accurately represent the current distribution of population (WorldPop 2025).

56. When pre-existing frames are insufficient, researchers may manually create new ones either using a simplistic method (e.g., by dividing maps into equally sized grid squares to cover the survey area comprehensively) or more computational methods that construct PSUs with roughly equal population based on gridded population datasets (Thomson 2022). However, it is important to note that not all area frames rely on gridded population data; most are derived directly from administrative divisions or census enumeration areas.

57. One of the primary advantages of using area frames is their applicability in situations where up-to-date or comprehensive lists of dwellings or households do not exist. Researchers can create PSU-specific sampling frames through field enumeration (ICF International 2012). This approach ensures high coverage and reduces the risk of missing target population segments.

58. However, area frames also have notable disadvantages. The household listing process is labour-intensive and time-consuming, requiring enumerators to physically visit each selected area to list all dwellings or households, which increases survey costs and adds logistical complexity. Additionally, there is a risk of selection bias if the same fieldworkers conducting the enumeration are responsible for selecting dwellings for the survey. To mitigate this issue, the enumeration and selection processes should be performed independently, with dwelling selection ideally conducted centrally rather than in the field.

Listing vs. random route techniques

59. An important consideration is whether to implement listing techniques or random route methods within selected PSUs. A PSU refers to the first unit selected in a multi-stage sampling design, often a geographical area such as a district, block, or census area, which serves as the basis for further sampling stages.

60. Field listing involves systematically enumerating all dwellings in a selected area to create a comprehensive sampling frame from which dwellings or households can be randomly selected. In contrast, random route techniques involve interviewers selecting dwellings based on a walking path through the PSU or a set of rules during fieldwork.

61. While random route methods may seem practical, they cannot assign known probabilities of selection to each dwelling, violating a core principle of probability sampling. Without known selection probabilities, it becomes impossible to calculate accurate weights or inferential statistics which compromises the validity of survey results. Despite these limitations, random route methods are still widely used in some survey

practices. Therefore, it is essential to explicitly highlight their limitations and emphasize the superiority of field listing for maintaining methodological rigor.

Integration with previous surveys or census data

62. Area frames may also be derived from previous surveys or census data. When a previous survey has already generated a robust sampling frame through field enumeration or mapping processes, this frame can be reused or updated for subsequent surveys. This approach can reduce costs and time requirements while leveraging existing resources. However, researchers must carefully validate and update such frames to account for demographic changes or other factors that may have affected the original frame's accuracy.

63. In summary, area frames with field enumeration are valuable for survey sampling, particularly in contexts where other sampling frames are unavailable, incomplete, or outdated. They offer comprehensive coverage through systematic listing, making them especially useful in regions with dynamic population changes or informal housing arrangements. However, their implementation requires careful attention to enumeration practices, independence between listing and selection processes, and rejection of random route techniques in favour of probabilistic approaches.

5.2.4. Common sources for sampling frames

64. Access to a reliable and comprehensive sampling frame is essential for designing an efficient sample. Several common sources are typically used to construct sampling frames, each offering distinct advantages and limitations. The most frequently utilized sources are:

[1] Administrative or statistical registers
[2] Census data
[3] Previous surveys
[4] Telephone-based frames

5.2.4.1. Administrative or statistical registers

65. Administrative or statistical registers are valuable sources of sampling frames due to their detailed and frequently updated information on households and individuals. NSOs (National Statistical Offices) often maintain these registers through regular administrative processes, such as tax filings, social security systems, or health insurance records. For example, countries like Sweden, Denmark, and the Netherlands rely heavily on administrative registers maintained by their NSOs for population-based sampling (Jabkowski and Kołczyńska 2020). These registers offer high accuracy, frequent updates, and comprehensive coverage, enabling efficient stratification and sampling across various demographic and geographic domains. However, limitations arise due to data privacy laws, which can restrict access, and the periodicity of updates may vary depending on the specific register.

5.2.4.2. Census

66. A population census is one of the most complete and widely used sources for constructing sampling frames. It typically includes data on geographic location, household size, and detailed demographic characteristics. While these elements enhance sample creation efficiency, they are not strictly essential for representativeness. NSOs, such as the U.S. Census Bureau or the Office for National Statistics in the United Kingdom, use census data to support sampling efforts for official surveys and other statistical programs (Bethlehem 2002).

67. While census data provide full population coverage and detailed demographic insights, they are generally conducted once every ten years, which means the data can become quickly outdated. This issue poses challenges in rapidly changing populations, such as urban centres or regions with high migration

rates. For example, many African countries, like Kenya and Nigeria, rely heavily on census data to construct sampling frames for national household surveys, even though data timeliness remains a recurring challenge.

5.2.4.3. Previous surveys

68. Previous surveys are valuable sources for sampling frames, mainly when designed to represent the population. NSOs often reuse sampling frames from earlier surveys or adopt multi-stage designs, where previous samples serve as a foundation for selecting new ones (Lepkowski and Couper 2001). For instance, in Mexico, the National Institute of Statistics and Geography (INEGI) frequently uses sampling frames from previous household surveys to design new ones, enabling cost-effective and efficient sampling (INEGI 2020).

69. A typical application of previous surveys is in two-phase or double sampling, where an initial survey collects basic information, and a subset of respondents is resampled for more detailed data collection. This technique is beneficial when the original sample is large, and a follow-up study with a smaller subset of respondents is required. However, the usefulness of previous surveys depends on their age and the degree of demographic or structural change in the population. Additionally, any biases or response errors from the original survey may carry over into the new sampling frame.

5.2.4.4. Telephone-based frames

70. The COVID-19 pandemic accelerated the adoption of telephone and online surveys as primary data collection methods (Maslovskaya, Struminskaya, and Durrant 2022). This shift was driven by mobility restrictions and lockdowns, which made in-person interviews impractical and, in many cases, impossible. As a result, telephone-based surveys became an essential tool for gathering timely information, especially in scenarios requiring real-time monitoring, such as health crises or economic disruptions. Mobile phone-based sampling frames provided an agile and cost-effective method for real-time data collection, offering a significant advantage during crises (Himelein et al. 2020). Despite severe logistical and operational challenges, they enabled NSOs to continue their survey operations.

71. However, telephone-based sampling frames face inherent limitations, primarily related to coverage, accessibility, and geographic representativeness. Despite these challenges, they offer several advantages, including cost-effectiveness, as telephone surveys generally have lower operational costs than face-to-face surveys, reducing transportation and field staff logistics expenses. They also provide speed, allowing faster data collection, which enables quicker responses to emergencies or policy changes. Additionally, surveys can be conducted across wide geographic areas without requiring physical presence, increasing their reach. Finally, telephone-based sampling frames demonstrate adaptability by integrating technologies, such as CATI, to streamline data collection and reduce interviewer errors.

72. One of the primary challenges in using telephone-based frames is linking mobile phone numbers to specific geographic areas, as mobile numbers are not inherently tied to a fixed geographic location due to number portability and user mobility. Strategies for overcoming geographic linking challenges include several approaches:

[1] Area codes: While area codes can approximate geographic locations, their utility is limited due to number portability and user mobility (Brick et al. 2007).

[2] Enriched databases: Some NSOs use commercial or government databases that link phone numbers to geographic units, such as postal codes. However, access to these databases can be costly and subject to negotiation (Kalton 2020).

[3] Previous surveys or censuses: Collecting phone numbers in face-to-face surveys or censuses provides a reliable method for linking mobile numbers to geographic areas. This strategy is often implemented as two-phase sampling, where auxiliary data from earlier surveys help adjust for nonresponse bias in phone surveys (Peytchev et al. 2020).

[4] Predictive models: When direct geographic data is unavailable, predictive models can be built using mobile usage patterns, demographic characteristics, and contextual factors.

[5] Hybrid approaches: Combining telephone-based frames with other sampling methods, such as face-to-face enumeration or online panels, can help improve coverage and reduce selection biases.

73. Additionally, telephone-based frames involve several key operational considerations, including:

[1] Bias and representation: Mobile phone ownership is not evenly distributed across demographic groups. Populations in rural areas, lower-income households, or older age groups may have lower access to mobile phones, leading to coverage bias.

[2] Response rates: Nonresponse rates are often higher in telephone surveys than face-to-face interviews. Factors such as call fatigue, distrust in unsolicited calls, and spam filters contribute to this issue.

[3] Data privacy and consent: Stringent privacy regulations may limit access to mobile phone databases and complicate the informed consent process for participants.

[4] Technological infrastructure: Reliable telecommunication networks are essential for the effective implementation of phone surveys, especially in remote or underserved areas.

[5] Questionnaire design: Survey questions need to be clear and concise due to the limited attention span of respondents over phone calls.

74. The application of telephone-based sampling frames in low- and middle-income countries (LMICs) has proven to be valuable for conducting surveys in challenging environments. However, these contexts often face unique barriers:

[1] Low mobile phone penetration: Despite significant growth in mobile phone ownership, rural and low-income populations remain underrepresented, introducing coverage bias (James 2021).

[2] Lack of reliable administrative databases: Many NSOs lack updated and complete phone number databases, complicating the construction of representative frames.

[3] High access costs: Commercial databases are often prohibitively expensive, requiring substantial financial investments (Kalton 2020).

[4] Regulatory limitations: Privacy laws and data protection regulations may restrict access to phone number databases, necessitating innovative sampling strategies (Groves 2011).

[5] Digital divide: Uneven access to smartphones and reliable mobile networks creates inequities in participation, compromising the representativeness of the sample.

75. Emerging solutions in LMICs focus on innovative strategies to overcome these challenges:

- Partnerships with telecom providers: Collaborative agreements with telecommunications companies can improve data accessibility and reduce costs.

- Integration of digital tools: Mobile survey apps and SMS-based questionnaires can enhance participation and reduce operational challenges.
- Community-based campaigns: Raising awareness about phone survey participation can improve trust and response rates in marginalized communities.

76. In summary, telephone-based frames provide an overview of their strengths and challenges:

- **Advantages:** They are cost-effective, allow for rapid implementation, are scalable across large geographic areas, and are adaptable to technology.
- **Challenges:** They require geographic linkage, leading to higher coverage bias, lower response rates, regulatory constraints, and reliance on reliable telecom infrastructure.
- **Solutions:** Hybrid approaches are welcome, with the implementation of predictive models, collaboration with telecom providers, use of enriched databases, and two-phase sampling.

77. In conclusion, telephone-based sampling frames represent an essential tool for modern data collection, especially in contexts requiring rapid response and wide coverage. However, their effectiveness depends on addressing geographic linkage, coverage bias, and operational limitations through innovative methodologies and strategic partnerships.

78. In summary, the choice of sampling frame source depends on the context, resources, and specific survey objectives.

- Administrative registers offer accuracy and coverage but may face privacy restrictions.
- Census data provide comprehensive demographic insights but risk becoming quickly outdated.
- Previous surveys offer cost-efficient solutions but depend on data timeliness and quality.
- Telephone-based frames are agile and cost-effective, but geographic linking and regulatory issues remain critical challenges.

Each source brings distinct advantages and limitations, and researchers must carefully evaluate their suitability for each survey project.

5.2.5. Choosing a frame

79. When selecting a sampling frame for household or individual surveys, it is essential to consider various factors influencing survey estimates' quality. The choice of a sampling frame can directly affect the representativeness of the sample and the validity of the survey results. Key factors to consider include:

- Mode of data collection (e.g., *PAPI, CAPI, CATI, CAWI*) (See Chapter 6)
- Budget constraints
- Available information for sampling design
- Precision requirements of the survey

These factors must be evaluated at the population level and across different estimation domains to ensure the survey meets its accuracy targets for various subpopulations of interest.

5.2.5.1. Coverage of the target population

80. Ensuring comprehensive coverage of the target population is fundamental to achieving valid and reliable survey results (S. L. Lohr 2008). A robust sampling frame should include every unit (household or individual) within the target population, minimizing exclusions that could introduce bias, which is particularly important when certain subpopulations (e.g., low-income households, specific ethnic groups, or geographically remote areas) are critical for policy decisions or analytical purposes. Neglecting these subgroups can result in skewed results, underrepresentation, and an inability to draw valid inferences about the broader population or the key subgroups.

81. An adequately designed sampling frame must list all units within the target population to support valid statistical inferences. A census-based frame is often utilized for household surveys, as it generally provides exhaustive coverage of all households in a given geographic region.

82. However, even a census frame may suffer from undercoverage, especially in dynamic environments where populations are mobile, or housing developments are rapidly evolving. Any gaps in coverage can result in biased estimates and hinder the ability to generalize findings; for example:

- A census-based frame might miss informal settlements or newly constructed housing in a rapidly urbanizing area.
- Lower-income or migrant households, which may exhibit distinct characteristics (e.g., employment rates and access to services), are often at risk of exclusion.

83. These groups may exhibit distinct characteristics (e.g., employment rates and access to services) crucial for valid inferences about the population. Omitting them compromises the inclusivity of the survey and risks introducing bias in the results. These omissions compromise survey inclusivity and increase the risk of introducing bias into the results.

84. When using dwelling frames, coverage may suffer if temporary housing units or informal dwellings are not included. In area frames, coverage relies heavily on the accuracy of geographic boundaries and the thoroughness of field enumeration. For telephone-based frames, challenges include number portability, incomplete databases, and uneven mobile phone penetration, all of which can limit coverage. Each frame type has unique coverage strengths and limitations, which must be carefully considered in the sampling design process.

85. Ensuring adequate representation of subpopulations is equally critical. A robust sampling frame should accurately capture key demographic and geographic subgroups, including:

- Age groups
- Ethnic minorities
- All socioeconomic levels
- Urban vs. rural populations

86. **Key considerations for inclusivity** include:

[1] Does the frame sufficiently represent marginalized or hard-to-reach populations, such as ethnic minorities, rural households, or lower-income groups?

[2] Is there a risk of exclusion of key subgroups due to administrative, logistical, or methodological constraints?

[3] How will undercoverage affect the precision and validity of survey inferences for the overall population and specific subgroups?

87. If key subgroups are underrepresented or excluded, the ability to make valid inferences about those groups—and the population as a whole—will be compromised.

Strategies for improving coverage

88. To maintain the quality and validity of survey inferences, researchers may need to supplement or adjust the primary sampling frame. **Key Strategies** include:

[1] Dual-frame sampling: Combining two frames (e.g., a census-based frame and an administrative register) can help address coverage gaps, but it simultaneously introduces challenges in covering some parts of the population units by two frames (see Section 5.2.8). For example, a household survey based on census data might underrepresent migrant populations and incorporating administrative records or community lists can improve coverage.

[2] Supplemental lists: Using localized lists from community organizations or administrative databases can enhance the representation of undercovered groups.

[3] Field enumeration: In area-based frames, careful field listing and enumeration can mitigate coverage biases.

89. Each approach must be carefully evaluated and tailored to the specific characteristics of the target population and survey objectives.

Evaluating coverage bias and its impact on inference

90. Coverage bias occurs when systematic exclusions of specific units from the sampling frame distort survey results. This bias poses a significant threat to the validity of survey inferences and must be addressed before sample selection.

91. Key steps to evaluate coverage bias include:

[1] Benchmarking: Compare the sampling frame's characteristics with known population benchmarks (e.g., national census data administrative records).

[2] Sensitivity analysis: Conduct sensitivity tests to estimate the impact of undercoverage on key survey indicators.

[3] Transparency: Document potential sources of bias and adjust survey weights accordingly.

92. When using dual-frame sampling or integrating alternative data sources, evaluating and documenting potential overlaps is essential to avoid duplication or misrepresentation.

5.2.5.2. Availability of contact information

93. The accuracy and completeness of contact information in sampling frames are essential for ensuring effective data collection and minimizing nonresponse bias. Depending on the data collection mode chosen – whether CATI, CAWI, or CAPI - the type of contact information required varies significantly.

- In CATI, having current and functional phone numbers is fundamental.
- In CAWI, while valid email addresses are often used, postal letters containing survey access credentials can also serve as an initial point of contact.
- In CAPI, accurate physical addresses are crucial for locating and reaching respondents effectively.

94. Contact information for sampling frames can be obtained from various sources: administrative records (e.g., tax registries, health insurance systems), telephone directories, commercial databases, census records, and previous survey phases where respondents have voluntarily provided contact details. Each source, however, has limitations that must be carefully assessed:

- Duplication: Repeated entries of the same contact information.
- Outdated entries: Invalid or obsolete phone numbers, email addresses, or physical addresses.
- Insufficient coverage: Exclusion of specific subpopulations or regions.

These issues can introduce systematic errors and reduce the effectiveness of survey operations, ultimately affecting the reliability of the results.

95. Assessing the quality and coverage of available contact information before finalizing the sampling design is essential. Accurate contact details determine the feasibility of reaching selected respondents across different modes, the efficiency of fieldwork operations, and the precision of survey estimates in reducing nonresponse bias. Research highlights that improvements in the quality of sampling frames (especially in contact information) can mitigate nonresponse risks and enhance survey representativeness. For example:

- Up-to-date phone numbers are critical for reducing the number of unanswered calls in telephone surveys.
- Valid email addresses or postal contact details improve initial outreach in web-based surveys.
- Accurate physical addresses ensure the efficient location of respondents in face-to-face interviews.

Tailoring contact information to data collection modes

96. Optimizing sampling frames requires aligning the type and quality of contact information with the specific requirements of each data collection mode.

- In telephone surveys (CATI), respondents without valid phone numbers are unreachable, limiting sample coverage.

- Email addresses in web surveys (CAWI) may not always be available or valid. However, alternative approaches, such as sending letters with survey credentials via postal services, can improve response rates.

- In face-to-face surveys (CAPI), outdated or imprecise physical addresses increase fieldwork costs and reduce operational efficiency.

97. Mixed-mode surveys, which combine different data collection techniques, highlight the importance of maintaining contact information across multiple formats (e.g., phone numbers, email addresses, postal addresses). Studies emphasize that the effectiveness of mixed-mode designs often depends on the consistency and completeness of contact information across modes.

Strategies to improve contact information quality

98. Improving the accuracy and completeness of contact information is a continuous process that requires:

[1] Validation and cleaning of databases: Regularly updating and cross-referencing contact details from multiple sources.

[2] Supplemental data collection: Incorporating contact details gathered during previous survey phases.

[3] Leveraging technology: Using automated validation tools to detect and correct errors in contact information.

[4] Customized outreach approaches: Adapting communication strategies (e.g., SMS reminders, postal invitations) based on available contact modes.

These strategies can significantly reduce nonresponse rates and improve survey operations' overall efficiency and reliability.

99. In summary, the availability and accuracy of contact information are central to the success of any survey operation, with its suitability depending on the chosen data collection mode and the survey context. In telephone surveys (CATI), functional phone numbers are essential for reaching respondents effectively. In web surveys (CAWI), valid email addresses or postal contact details can be reliable contact points to ensure participation. For face-to-face surveys (CAPI), accurate physical addresses are indispensable for locating respondents efficiently. Investing in the assessment, validation, and continuous improvement of contact information is crucial for reducing nonresponse bias and ensuring that survey findings remain representative and reliable.

5.2.5.3. Availability of auxiliary information to increase the sampling design's efficiency

100. One way to improve efficiency is by incorporating auxiliary information - additional data correlated with the primary survey variables - into the sampling design. The efficiency of a sampling design refers to its ability to produce precise and reliable estimates using the smallest possible sample size or resources.

101. Auxiliary information can be used to define strata or assign different selection probabilities during sampling, allowing researchers to focus resources on population subgroups with higher variability or lower response rates (Kreuter 2013). This approach ensures that the sample better reflects the population's key characteristics, improving the precision of estimates.

102. Ideally, auxiliary information is available at the micro-level, meaning that each unit in the sampling frame (e.g., each household or individual) is associated with specific values for these auxiliary variables (Bethlehem 2002). Commonly used auxiliary variables include education level, age, sex, household size, and geographic region. These variables are frequently available in national sampling frames and are valuable for improving sampling designs.

103. Auxiliary data also plays a key role in addressing nonresponse bias. When nonresponse is assumed to be Missing at Random (MAR) – a statistical assumption where the probability of response can be partially explained by auxiliary variables – it is possible to estimate response propensities using parametric or non-parametric algorithms (Little and Rubin 2019). These algorithms predict the likelihood of a selected unit participating in the survey based on the auxiliary variables, allowing researchers to adjust sampling weights and reduce bias.

104. Any remaining bias in modelling response propensities can be further reduced by incorporating auxiliary information into the estimation process. Among the most common techniques for integrating auxiliary information are regression estimation and calibration estimation.

- Regression estimator: This approach uses auxiliary variables in a regression model to predict survey outcomes. The predicted values are then combined with the observed survey data to

produce more accurate estimates. Regression estimation is particularly effective when the auxiliary variables have a strong linear relationship with the target variables.

- Calibration estimator: Calibration adjusts the sampling weights so that the weighted sums of the auxiliary variables in the sample match their known population totals. This process reduces the variability of the estimates, leading to lower standard errors compared to simpler estimators such as the *Horvitz-Thompson* estimator. Calibration is widely used in practice due to its flexibility and robustness in handling auxiliary data. For example, variables such as education level, age, sex, household size, or geographic region—often available in administrative registers or census data—can significantly improve the precision of survey estimates when used in calibration techniques. These adjustments ensure that the sample more accurately reflects the population structure, enhancing the overall reliability of the survey results (See Chapter 8).

5.2.6. Examples and strategies to effectively address issues related to imperfect sampling frames

5.2.6.1. Undercoverage

105. Undercoverage occurs when specific population segments are excluded from the sampling frame. For example, a census frame might fail to include newly constructed housing units or transient populations, such as migrant workers or refugees. In low- and middle-income countries, informal settlements or rapidly expanding peri-urban areas are often excluded due to outdated or incomplete data.

106. Strategies to address undercoverage include:

- Field listing operations: Conduct targeted field listing operations to update the frame, particularly in areas with frequent housing changes. This approach is especially relevant in informal settlements, where reliable maps or formal records may not exist.
- Supplementary data sources: Combine the primary sampling frame with alternative data sources, such as utility records, tax databases, or community-maintained records, to fill gaps in coverage.
- Community engagement: Collaborate with local community leaders or organizations to identify and map areas excluded from official records.

5.2.6.2. Overcoverage

107. Overcoverage occurs when the sampling frame includes units that no longer exist or are irrelevant, such as demolished buildings or unoccupied housing units resulting from migration or abandonment. These inaccuracies can inflate survey costs and affect data quality if not addressed.

108. Strategies to address over coverage include:

- Verification of units: Before data collection begins, verify the existence of units included in the frame using tools such as satellite imagery or targeted field visits.
- Pre-survey visits: Conduct pre-survey visits to confirm whether housing units are occupied and eligible for inclusion in the sample. This method is particularly cost-effective in areas with high housing turnover rates.
- Systematic exclusion rules: Establish clear rules and protocols to identify and exclude ineligible units systematically during data collection.

5.2.6.3. Missing or out-of-date contact information

109. Contact information, including addresses or phone numbers, may often be incomplete or outdated, creating barriers to reaching sampled units. In many low- and middle-income countries, the absence of comprehensive address systems, especially in rural areas or informal settlements, exacerbates this issue.

110. Strategies to address missing or out-of-date contact information include:

- Multiple data sources: Cross-reference multiple data sources (e.g., administrative records, previous surveys, commercial databases) to update and validate contact information.
- Local key informants: Involve community leaders or local key informants to assist in locating households without formal addresses. Their knowledge can significantly enhance the efficiency of outreach efforts.
- Mixed contact modes: Utilize a combination of contact methods, including in-person visits, phone calls, SMS, or email, to maximize the likelihood of reaching sampled households.

111. Addressing imperfections in sampling frames—whether under coverage, over coverage, or outdated contact information—is essential for ensuring the accuracy and reliability of survey data. Strategies such as field listing operations, supplementary data integration, community engagement, systematic verification, and mixed-mode contact approaches can significantly improve sampling frames' coverage, validity, and efficiency across different survey contexts.

5.2.6.4. Inaccurate auxiliary variables

112. Auxiliary variables, such as income or education level, are often used for stratification and sample allocation. However, when these variables are outdated or inconsistently available, they can reduce the efficiency and accuracy of the sampling design. This issue typically arises when the source of the sampling frame is outdated or when auxiliary variables are only available for some, but not all, units in the frame. The following strategies can be used to address inaccurate auxiliary variables:

- Regular updates to auxiliary data: Integrate routine updates to auxiliary data into ongoing statistical operations to ensure the sampling frame remains current and reliable. These updates can be achieved through administrative data integration or periodic revisions of the sampling frame.
- Simplified stratification variables: When detailed auxiliary data are unavailable or unreliable, stratifying by basic and stable characteristics, such as urban/rural classification or geographic region, can be an effective alternative to improve sample design.
- Imputation for missing auxiliary data: Apply statistical imputation techniques to estimate missing values for auxiliary variables using available data from similar units. This approach can reduce bias and maintain the utility of auxiliary variables in the sampling design.

5.2.7. Multiple frames

113. In household and individual surveys conducted by NSOs, achieving comprehensive population coverage often requires the integration of multiple sampling frames. This approach is particularly relevant in contexts where a single frame is insufficient to represent the entire population adequately.

114. A typical example involves combining an administrative data-based population registry with an area sampling frame. Government agencies often maintain population registries compiled from civil registries,

social security records, or tax databases. These registries are commonly used in high-income countries, such as Sweden, Denmark, or the Netherlands, where highly developed administrative data systems are regularly updated. However, in some low- and middle-income countries, population registries may be less reliable or entirely absent. Instead, household listing exercises or census data often serve as the foundation for sampling. In such cases, an area sampling frame becomes essential to complement or fill gaps left by administrative data sources. For example, in informal settlements, administrative records, such as municipal registries, social security databases, or electoral rolls—often fail to capture specific populations due to outdated records, incomplete registration processes, or restricted geographic coverage. This under-coverage can lead to biased estimates if these populations are excluded from the survey sample. Surveys can ensure more comprehensive coverage of these hard-to-reach populations by integrating an area sampling frame, which relies on geographical enumeration units and often includes a field-listing operation.

115. While integrating multiple sampling frames enhances population coverage and improves the inclusivity of survey designs, it also introduces operational and statistical challenges that must be carefully managed. One of the most common issues is overlapping coverage, where specific units, such as households or individuals, may appear in more than one frame, leading to duplication and potentially inflating the representation of those units. Another key challenge is differential selection probabilities, where the likelihood of selection varies across frames, creating inconsistencies in survey weights and increasing the risk of biased estimates. If these challenges are not adequately addressed through appropriate statistical adjustments, they can significantly compromise the accuracy and validity of survey results.

116. To address the operational and statistical challenges associated with multiple-frame sampling, several statistical techniques have been developed to ensure the accuracy and reliability of survey estimates. One widely used approach is the Dual-Frame Estimator, where the Horvitz-Thompson estimator, designed originally for single-frame sampling, has been extended to handle dual-frame contexts. This adaptation assigns specific selection probabilities to units in overlapping areas, effectively preventing duplication and ensuring proper representation (S. Lohr and Rao 2006).

117. Another essential technique is Regression and Calibration Estimators, which adjust sampling weights to harmonize the frames and correct discrepancies arising from overlapping coverage or under-representation. Calibration ensures that units in different frames contribute appropriately to the final estimates, thereby reducing variance and minimizing bias (Skinner 1991). For example, combining census data with an area sampling frame in a national household survey requires calibration techniques to guarantee that households appearing in both sources are not counted twice and that no eligible units are inadvertently omitted. These statistical approaches are essential for managing the complexities of multiple-frame sampling while maintaining the integrity and representativeness of survey results.

118. Calibration techniques play a crucial role in addressing discrepancies between multiple frames, such as over-coverage or under-coverage, by adjusting sampling weights. These adjustments ensure that the final survey estimates accurately represent the population composition, minimizing bias and improving precision. For example, if administrative records tend to over-represent urban areas while an area sampling frame captures more remote rural populations, calibration can balance these differences to produce unbiased estimates. Similarly, if some households appear in both frames, calibration ensures that these units are weighted correctly, preventing duplication and distortion of survey results. Through these adjustments, calibration harmonizes the contributions of different frames, improving the reliability and validity of survey findings.

119. In summary, integrating multiple sampling frames—such as administrative records and area sampling frames—is a powerful strategy for addressing under-coverage and enhancing the

representativeness of household and individual surveys. However, this approach also introduces significant operational and statistical challenges. Techniques such as dual-frame estimators, regression estimators, and calibration methods provide robust solutions for managing overlaps, correcting selection biases, and harmonizing data across different frames.

120. By carefully evaluating the strengths and limitations of each frame and applying appropriate statistical adjustments, researchers can significantly improve the accuracy, reliability, and inclusivity of survey estimates, ensuring that the results are methodologically sound and reflective of the target population.

5.2.8. Documenting frame information

121. Sharing relevant metadata and information about the sampling frame with data users is essential to ensure transparency and enhance the credibility of statistical outputs. Metadata refers to detailed documentation describing the characteristics, methodology, and quality of the sampling frame, and it serves as a critical resource for enabling users to make valid inferences based on survey data. Proper documentation allows users to understand the sampling frame's strengths, limitations, and coverage, facilitating informed interpretation of survey results. These metadata should be included in the Technical Report and other data dissemination products, as emphasized in Chapters 1 and 10.

122. Below are key recommendations on what frame metadata and information should be shared with users.

5.2.8.1. Frame exclusions and coverage information

123. Providing precise details about the coverage of the sampling frame and any exclusions is crucial for ensuring transparency and helping users assess the representativeness of the survey. Coverage information should specify the frame's geographic, demographic, or institutional boundaries and clarify whether the frame fully encompasses the target population or has known limitations.

124. Frame exclusions refer to units that are part of the target population but were intentionally or unintentionally omitted. For instance, intentional exclusions might include remote rural areas or regions with limited accessibility, often omitted due to budget constraints or operational difficulties. Sharing this information enables data users to quantify potential biases in the survey estimates resulting from these exclusions.

125. For example, the UNECE guidelines on using administrative data emphasize the importance of documenting coverage issues and their potential impact on survey results (United Nations Economic Commission for Europe 2019). By providing detailed coverage metadata, data users can better evaluate the strengths and weaknesses of the sampling frame and adjust their analyses accordingly.

5.2.8.2. Sampling frame methodology

126. It is essential to thoroughly document the methodology used to construct the sampling frame. Metadata should include details about:

- Data sources: Whether the frame was built using administrative records, survey data, or multiple sources.
- Integration processes: How different sources were combined and harmonized to address discrepancies or overlaps.
- Data cleaning methods: Procedures for removing duplicates, managing over-coverage, and ensuring consistent classification across sources.

127. For example, suppose a sampling frame for a national household survey integrates census data and tax records. In that case, it is important to explain how inconsistencies between sources were resolved to ensure that households appearing in multiple registries were not double-counted (a problem known as multiplicity).

128. Providing these methodological details allows users to understand the assumptions, processes, and potential biases in the sampling frame, improving their ability to interpret survey results accurately.

5.2.8.3. Update frequency

129. The timeliness of the sampling frame (e.g., how frequently it is updated and how recent it is relative to the data collection period) is a key factor influencing the quality of survey estimates. Metadata should specify:

- Update intervals for different variables in the frame.
- Stability or variability of frame variables over time.

130. For example, structural variables (e.g., household size or geographic region) tend to remain stable over time, while dynamic variables (e.g., migration rates, employment status) may change rapidly.

131. Sharing update frequency metadata helps users assess whether the recency of the frame aligns with the timing of data collection and the dynamic nature of the target population. Guidance from Eurostat underscores the importance of evaluating the impact of frame updates on survey design and results (European Commission. Statistical Office of the European Union. 2015).

5.2.8.4. Frame quality indicators

132. Quality indicators provide valuable insights into the reliability and accuracy of the sampling frame. These indicators should include:

- Non-contact rate: Result from deficiencies in the sampling frame, particularly related to inaccurate contact information.
- Under-coverage and over-coverage rates: Metrics highlighting gaps or excesses in coverage.

133. These quality indicators should be computed and reported globally and for different domains of analysis (e.g., regions, urban/rural areas) typically used in surveys derived from the same frame. Publishing these indicators enables data users to assess potential sources of error or bias and evaluate the reliability of survey results across different domains of interest.

134. In summary, documenting and sharing metadata about the sampling frame is essential for ensuring survey data's transparency, credibility, and usability. Detailed information on coverage, exclusions, methodology, update frequency, and quality indicators must be included in the Technical Report and other data dissemination products. These metadata not only enable users to understand the limitations and strengths of the sampling frame but also empower them to interpret survey results more accurately and make informed decisions based on the data.

5.3. Sampling

5.3.1. Introduction

135. The United Nations (UN) has provided comprehensive guidance on sampling for household surveys through a series of valuable documents over the years (United Nations 2008a; 2008b; 1986; 1984; 2005). The fundamental principles remain unchanged, and the statistical theory and practical recommendations

presented in these documents are robust and relevant. Current household survey practices align closely with the methods described in these foundational resources.

136. This section serves as an overview of key sampling concepts, with references to more extensive documents for detailed coverage of specific topics. Readers are encouraged to consult earlier UN guidance and other authoritative resources on sampling (Brick 2011; Jabkowski and Kołczyńska 2020; Kalton 2020; de Leeuw, Hox, and Dillman 2008; S. L. Lohr 2021a).

137. Although the core principles of sampling remain constant, there have been significant changes over time in how sampling is implemented, driven by technological advancements, shifts in survey participation trends, the evolution of sampling frames, and the increasing use of non-probability sampling methods. These changes can be summarized as follows:

[1] As seen in the previous sections of this chapter, technological advancements have transformed the way sampling frames are developed and utilized. Today, many sampling frames are georeferenced, including digitized household addresses or coordinates of PSU boundaries. These geo-coordinates enable precise planning of field operations using digital mapping tools and routing algorithms.

[2] There has been a notable decline in survey participation rates in many countries, with increasing portions of the population reluctant to respond (Jabkowski and Cichocki 2024). This reluctance poses logistical challenges, as interviewers may conduct fewer full-length interviews than planned. It also impacts survey precision, as nonresponse adjustments can increase the variability of survey weights (See Chapter 8 for more details). Addressing these issues requires advanced statistical methods to manage missing data effectively (See Chapter 7). Survey planners are encouraged to consult recent examples or case studies from surveys conducted in similar contexts to anticipate and mitigate these challenges.

[3] Non-probability sampling has become more common when probability sampling faces daunting operational, cost, or logistical constraints. While this Handbook acknowledges the relevance and growing use of non-probability sampling techniques and provides references and examples for their application, it recommends probability sampling for most government surveys. Probability sampling remains the gold standard due to its statistical rigor and ability to produce unbiased, representative results.

5.3.2. Goals of sampling design

138. Survey researchers rely on sample data to estimate results that would be obtained from a complete census if it were feasible to collect information or measurements from every individual in the population. The primary objective is to ensure that inferences drawn from the sample to the population are unbiased and precise. Additionally, surveys must be cost-effective and conducted within a reasonable timeframe. Nearly all decisions in designing a survey sample stem from balancing these goals: obtaining unbiased, precise, timely, and affordable estimates of population characteristics.

139. A sampling frame (see previous sections**)** plays a fundamental role in meeting these goals. The quality and completeness of the sampling frame directly influence the accuracy and reliability of survey estimates.

140. Unbiased means that the expected value of the survey estimate aligns with the true, unknown population parameter. Precise refers to sampling variability or the degree of deviation of the estimate from the unknown population parameter, remaining within an acceptable range for decision-making or action.

Timely and affordable are context-dependent and defined by the stakeholders commissioning and funding the survey.

141. Probability sampling remains the preferred approach for most applications to achieve these goals. A probability sample is one where 1) every member of the population under study has a non-zero probability of being selected and 2) it is possible to calculate the selection probabilities for those included in the sample.

142. The foundational premise of probability sampling ensures that conclusions drawn from survey data are:

[1] **Representative of the broader population:** The probability of selection provides a statistical link between individuals included in the sample and those who were not selected but could have been. This connection ensures that results are generalizable to the entire population.

[2] **Unbiased, or nearly so:** Some survey estimators are designed to produce results that, on average, are equally likely to be slightly above or below the unknown population parameter. Details regarding the bias and precision of specific estimators can be found in their computational documentation.

[3] **Precise:** Sampling variability—or the uncertainty in survey estimates resulting from random sample selection—is quantifiable. This variability, often referred to as **sampling error**, does not imply methodological mistakes but highlights that survey estimates may deviate from the true population parameter within predictable, probabilistic bounds.

5.3.3. Types of sampling designs

143. A sample design is a structured plan to select units from a sampling frame to a survey sample. This plan determines how sampling units are chosen, how sample size is allocated, and how estimation will be performed (See Chapter 8). A well-designed sample aims to balance representativeness, efficiency, and practicality while ensuring that survey results are unbiased, precise, timely, and affordable.

5.3.3.1. Complete enumeration / census

144. If feasible, stakeholders often prefer collecting data from every member of the population of interest (e.g., determining employment status for all adults or nutritional status for all children). However, conducting a full census is rarely practical due to resource constraints and logistical challenges. As a result, sample surveys are commonly used to estimate unknown population parameters.

145. A country's periodic census plays a central role in shaping available sampling frames. While some countries conduct censuses every ten years, others may do so every five years. However, not all countries can consistently achieve high-quality census operations due to financial, logistical, or political challenges. Census data is often used to refresh sampling frames, create updated lists of census enumeration areas, and enrich the information for describing sampling units. This updated data can include geographic boundaries, population counts, and auxiliary variables used for stratification and sampling design.

146. To transition to the next sections: Although a census provides comprehensive population data, it is often impractical for most statistical projects. Consequently, sample designs are employed to achieve representative and reliable estimates efficiently.

147. **Note***:* More detailed discussions on how census data refresh sampling frames are covered in Section 5.2 (Sampling Frames) of this chapter.

5.3.3.2. Simple random sampling (SRS)

148. In a *Simple Random Sample* (SRS), every possible sample of a given size has an equal probability of being selected. Consequently, every unit in the population has the same probability of inclusion in the sample.

149. A classic analogy is drawing names from a hat: every name (representing an individual or household) is written on a slip of paper, placed in a hat, mixed thoroughly, and a subset is selected randomly. Every possible subset of the same size has an equal chance of selection, making the sampling process fair and unbiased.

150. For estimation purposes, an SRS is often considered self-weighting, meaning that every selected unit represents an equal number of non-selected units. Because of its simplicity and mathematical clarity, SRS serves as a benchmark for evaluating the efficiency of more complex sampling designs.

5.3.3.3. Systematic sampling

151. *Systematic sampling* involves selecting units from an ordered list at regular intervals, known as the *sampling interval* until the desired sample size is achieved.

- The sampling interval is calculated by dividing the number of units in the sampling frame by the desired sample size.
- The first unit is selected randomly, and subsequent units are chosen by applying the sampling interval as an offset.

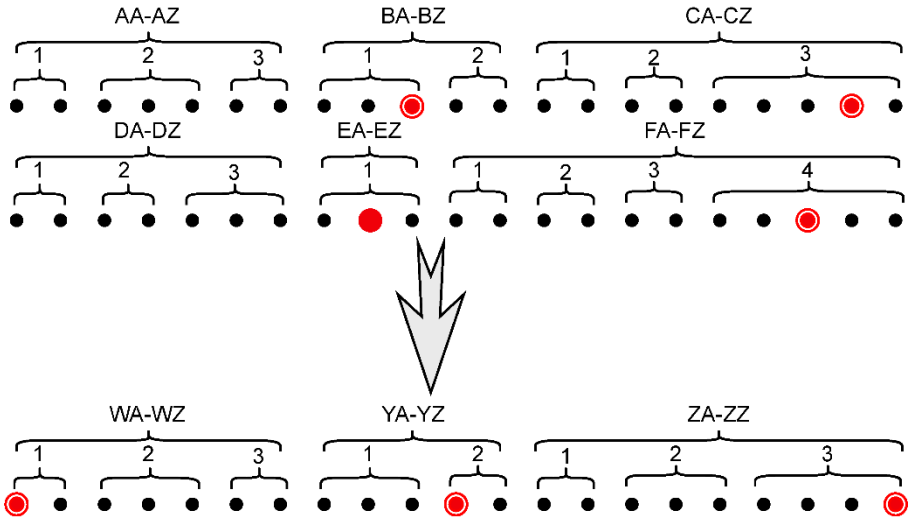
152. The order of the sampling frame impacts the efficiency of systematic sampling:

- If the frame is sorted by a variable related to the survey outcome, systematic sampling can yield a more statistically efficient sample than SRS.
- If the sorting introduces periodic patterns correlated with the survey outcome, it can reduce sampling efficiency.
- If the frame is sorted by geographic or demographic variables (e.g., urban/rural classification), systematic sampling can achieve implicit stratification, ensuring balanced representation across key subgroups.

153. **Figure 5.2** visualizes a systematic sampling design where frame units are sorted by two variables (variable 1 has string values, and variable 2 is numeric). Assume that the size of the population is 200, and the goal is to select a sample of 20 units (the interval for systematic sampling is then 10). The process starts by randomly selecting one unit for the sample (solid red dot in **Figure 5.2**), and then (starting from the selected one) the process involves systematically draw every 10th to the sample (red dot with a hollow border).

154. Systematic sampling is widely used because it is straightforward and often provides favourable sample allocation relative to the sampling frame's structure.

Figure 5.2. Illustration of systematic sampling design.



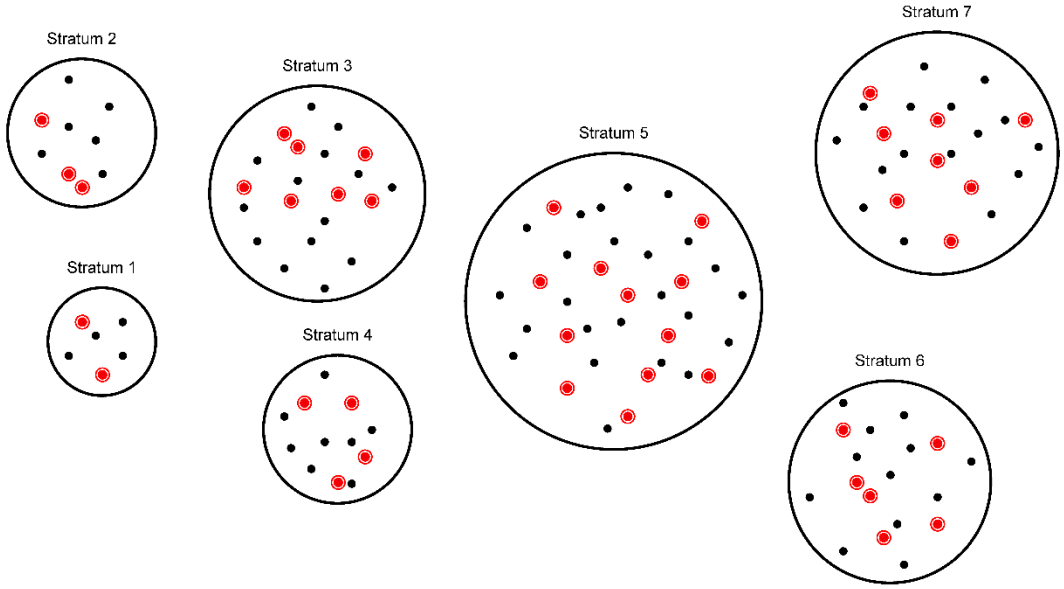
5.3.3.4. Stratified sampling

155. In *stratified sampling*, the population is divided into strata (subgroups) defined by shared characteristics, such as geographic location or demographics. The singular form of *strata* is *stratum.* These strata are mutually exclusive (no unit belongs to more than one stratum) and exhaustive (every unit belongs to one stratum). An independent sample is drawn from each stratum, and the results are combined to make inferences about the entire population.

156. **Figure 5.3** presents stratified sampling proportional to the strata size (see Section 5.3.6. Allocate PSUs across strata, for more details) with a sample of $n = 42$. The units (e.g., households or individuals) are grouped into seven strata, and then a random sample is independently selected from each stratum (the size of each stratum determines the number of selected units within the strata).

157. Whenever a stratum's members are more homogeneous in some way than would be true of a mix of people from different strata, a stratified sample design improves the precision of survey estimates when compared to a SRS. It is important to note, however, that stratification is always recommended even if it is not beneficial for precision as it allows for control of the distribution of a stratum variable in a selected sample.

Figure 5.3. Illustration of a stratified sampling design.



158. Advantages of stratified sampling include:

- Increases precision when members within a stratum are more homogeneous than across strata.
- Guarantees representation of all subgroups of interest, even if they are small.
- Allows different sampling plans (e.g., oversampling but with a cost of increased variance due to introducing variability in selection probabilities) for each stratum, which allows for efficient resource allocation.

159. Whenever possible, the survey's main estimation domains or areas should be defined as strata, allowing the sample size to be controlled and fixed if the sampling design permits it. For example, in a national household survey, geographic regions (e.g., provinces or districts) are often defined as strata to ensure reliable estimates at both regional and national levels. These domains, known as planned domains, are designed to minimize the increase in the standard errors of the estimates caused by the random variability in the domain's sample size. This approach is particularly important for surveys that require reliable estimates for specific subpopulations (Singh, Gambino, and Mantel 1994).

5.3.3.5. Clustered and multi-stage sampling

160. In cluster sampling, the population is divided into comparatively small groups, often based on geography. Clusters might be defined using boundaries of villages or townships or census enumeration areas. These clusters serve as the Primary Sampling Units (PSUs). A common design for a nationally representative household survey is to conduct a cluster survey within each stratum in the population and combine all the data for analysis.

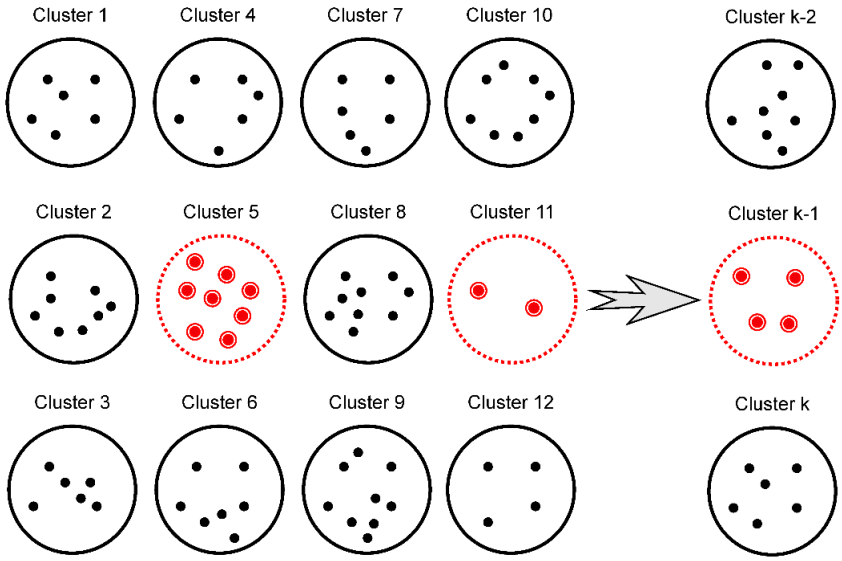
- Single-stage cluster sampling: Every household or individual within selected clusters is interviewed.
- Multi-stage cluster sampling: Selection occurs in multiple stages. For example:

- Stage 1: Select clusters (e.g., villages).
- Stage 2: Select households within selected clusters.
- Stage 3: Select individuals within households.

161. For example, in a single-stage cluster design, every person in every selected cluster is interviewed. In a multi-stage household survey, the sampling staff might select clusters and visit each to establish an up-to-date frame of households; then randomly select households, visiting each one to establish an up-to-date frame of adults; select one adult per household and conduct interviews with them.

162. **Figure 5.4** presents an example of cluster sampling, where the population is divided into *k* clusters (e.g., blocks). Three of these clusters are randomly selected, and all units (e.g., households) within the selected clusters are included in the sample.

Figure 5.4. Illustration of a one-stage cluster sampling.

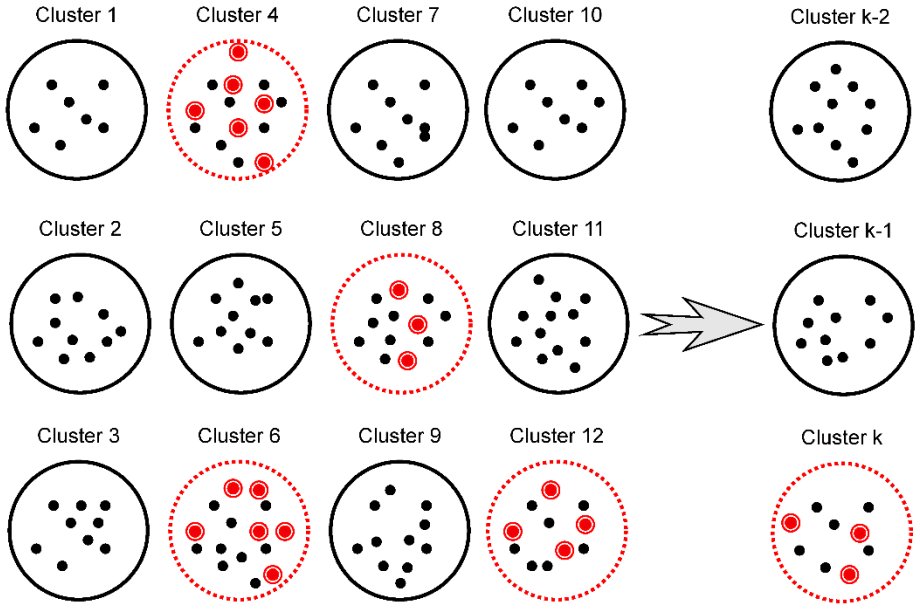


163. In face-to-face interviews with geographically dispersed populations, cluster sampling is generally more cost-effective than simple random sampling (SRS) because, after investing time and resources for an interviewer to travel to a PSU, it is more efficient to interview several nearby respondents before moving to a new location. However, respondents who live close to each other are often not statistically independent. Their access to services and information is likely to be similar, and their attitudes may also share similarities, shaped by regional influences or shared factors that influenced their choice of where to live.

164. In the previous example, it is important to note that in cluster sampling, the final sample size (e.g., households) is random (i.e., not fixed). This variability affects both the precision of the estimates, and the budget allocated for the survey. The issue becomes more significant when there is substantial variability in the sizes of the PSUs.

165. Expanding on the previous discussion, **Figure 5.5** illustrates an example of two-stage cluster sampling. The population is divided into *k* clusters (e.g., blocks), three of which are randomly selected. Within these selected clusters, some units (e.g., households) are randomly chosen to be included in the final sample with a probability proportional to the cluster size. When it comes to a sample drawn from the frame of households, it is more effective to select more households and one person within a household than to select fewer households but include all eligible household members.

Figure 5.5. Illustration of a two-stage cluster sampling.



166. Because of this reduced statistical independence, cluster sampling typically reduces the precision of survey estimates; the reduction is quantified with a factor often referred to as a "design effect" (*DEFF*). The extent of similarity among nearby respondents is quantified using an intracluster correlation coefficient (*ICC*), a statistical measure that indicates how strongly individuals within the same cluster resemble each other in relation to a specific survey variable.

167. An advantage of cluster sampling is to reduce **travel costs** in face-to-face surveys by concentrating interviews geographically. One disadvantage of cluster sampling is increased **design effects** due to **intra-cluster correlation**, where respondents within the same cluster may share similar characteristics. Further, cluster sampling requires adjustments in variance estimation using tools that account for the survey's design effect.

168. **Recommendations for Creating Primary Sampling Units** include:

- PSUs should not overlap and must cover the entire survey population.
- Verify that every household or unit in the target population is included in exactly one PSU.
- Design PSUs so that fieldwork effort per PSU is balanced.
- PSUs should have enough households to allow robust second-stage sampling. Avoid PSUs that are too small (insufficient households) or too large (logistically challenging). In densely populated areas, PSUs might correspond to smaller geographical units (e.g., city blocks).
- In sparsely populated areas, PSUs might represent larger regions (e.g., entire villages).
- If possible, stratify PSUs by key variables (e.g., urban/rural status).
- PSUs should aim to maximize homogeneity within clusters while maintaining variability across clusters.

- Use last census data, administrative records, satellite images, or gridded population datasets combined with GIS shapefiles to define PSUs.

5.3.3.6. Probability proportional to size (PPS) sampling

169. In *Probability Proportional to Size* (PPS) Sampling, each unit in the population is assigned a probability of selection proportional to its size. The "size," often referred to as the *Measure of Size* (MOS), typically refers to an auxiliary variable, such as the number of individuals, households, or other measurable attributes associated with each unit. PPS sampling ensures that larger units have a higher probability of being selected. Once the clusters are selected, a secondary sampling method, such as SRS, is typically used to select individual units (e.g., households or individuals) within each chosen cluster.

170. If the sampling frame is up-to-date and the MOS is accurate, then when PPS sampling is used in the first stage and a fixed sample size is selected in the second stage, the result is a self-weighted sample. In this scenario, each secondary unit has the same probability of inclusion, resulting in equal sampling weights based on the design. This characteristic simplifies the estimation process and reduces the complexity of weighting adjustments during data analysis.

Identifying certainty PSUs

171. The selection probability of PSU i under a PPS design without replacement is defined as $\pi_i = n \times p_i$, where n is the target number of PSUs and p_i represents the relative size of PSU i. In some cases, the size of certain PSUs is so large that they are selected into the sample with a probability of 1. This occurs when $n \times p_i$ exceeds 1, and such PSUs are said to be selected with certainty and sometimes called certainty PSUs. This situation commonly arises in large metropolitan areas with high population density.

172. Once certainty PSUs have been identified, the relative sizes of the remaining PSUs in the frame must be recalculated for the final sample selection, and the selection probabilities are defined as $\pi_i = (n - K) \times p_i$, where K is the number of certainty PSUs in the stratum. This procedure is repeated iteratively until there are no more PSUs where $n \times p_i \geq 1$.

173. Note that when certainties PSUs are present and the second-stage sampling assigns the same number of units (e.g., households) to each PSU, the resulting sample is not self-weighting. During analysis, certainty PSUs should not contribute to variability of stratum level estimates, so each is treated as an individual stratum (see Chapter 9).

5.3.3.7. Multi-phase sampling

174. A *multi-phase sample* involves conducting interviews on two or more separate occasions. The first phase is typically brief and serves to screen selected respondents for eligibility in later phases. For example, it might identify individuals with a rare condition that is not directly observable (e.g., children living in homes with lead paint on walls or window frames). In subsequent phases, individuals meeting the condition of interest are oversampled for more detailed interviews or resource-intensive measurement protocols.

175. This approach enhances cost efficiency by reducing the burden on all respondents during the initial phase and ensuring that costly equipment and specialized personnel are deployed only for respondents of interest.

176. *Two-phase sampling* differs from two-stage sampling because it does not satisfy the *invariance* assumption. In simple terms, this means "*investigators cannot just look at what came out in the first stage.*" In two-phase sampling, the results from the first phase can directly influence the sampling probabilities or

rates applied in the second phase. For instance, different sampling rates might be chosen for different clusters based on the results of the first phase, allowing for a more targeted allocation of resources.

177. For example, in a household survey conducted by a NSO, the first phase might involve a brief screening questionnaire applied to a sample of households across various geographic regions to identify those with elderly individuals living alone. In the second phase, households identified as having elderly individuals living alone could be oversampled for a more detailed follow-up interview focusing on topics such as access to healthcare, social support, and living conditions. In this scenario, the sampling rate in the second phase is adjusted based on the results of the first phase, violating the assumption of invariance and requiring appropriate weighting adjustments during the analysis to ensure unbiased estimates.

5.3.3.8. Key concepts of sampling design

178. The designs mentioned thus far are often combined. Below are key steps commonly found in a national household face-to-face survey:

[1] **Stratification by subnational states:** The sample might be stratified by subnational states or provinces to ensure representation across different regions.

[2] **Cluster sampling for cost-effectiveness:**

- Within each state, cluster sampling is often used to reduce travel costs and logistical complexity.
- Clusters might be selected systematically from a frame sorted by urban and rural clusters and further ordered by sub-sub-national domains (e.g., counties). The selection would often be with probability proportional to size.

[3] **Household listing and sampling within clusters:**

- Once clusters are selected, a list of households within each selected cluster might be listed or enumerated using a census approach.
- From this list, a simple SRS or systematic sample of households could be drawn for interviews.

[4] **Selection of individuals within households:** Inside each selected household, a random adult might be chosen using simple random sampling.

[5] **Additional stages in some designs:** In certain surveys, more stages may be added beyond those described above, depending on the survey's goals and complexity.

[6] **Complex sample design and variance estimation:**

- Any design that employs stages, strata, or clusters is typically referred to as a *complex sample design*.
- The term *complex design* implies that the variance of survey estimates should not be calculated as if the sample were from a SRS.
- Instead, specialized survey analysis software should be used to account for the design's structure and apply the appropriate adjustments (bonus or penalty) to the variance estimates.

[7] **Documentation in survey analysis software:**

- Survey analysis software typically includes clear documentation on how to define the sample design using syntax.

- It also provides guidelines on how to compare the variance from a complex sample design to what might have been observed with an SRS of the same sample size.

179. Several less common designs are also important for specific applications, described in Sections 5.3.3.9 through 5.3.3.12.

5.3.3.9. Indirect sampling

180. *Indirect sampling* is a method used when studying populations for which a direct sampling frame is unavailable, and it allows the inclusion of groups that are difficult to identify or access through traditional sampling methods. For example, consider a survey conducted by a NSO aiming to understand household expenditure patterns among families with irregular income sources.

181. In such cases, it can be efficient to interview a randomly selected group of respondents and ask about their knowledge of acquaintances who belong to the target population. Respondents might be asked if they know anyone in the group of interest, how many people they know, and to provide general information about the behaviour or characteristics of those acquaintances.

182. The objective of indirect sampling is not necessarily to conduct follow-up interviews with the individuals identified through respondents' networks but to gather information about the population of interest when no direct sampling frame exists.

183. This method is particularly useful when social stigma, privacy concerns, or logistical constraints make direct contact with the population of interest challenging. However, analysing data from indirect sampling requires specialized statistical techniques to account for potential biases. These biases can arise if the connections between respondents and subjects are not representative (e.g., a respondent is a social worker who knows an unusually high number of individuals from the target group) or if respondents' knowledge about their acquaintances is inaccurate.

184. Additionally, it may be relevant to mention non-probability methods, such as Respondent-Driven Sampling (RDS), which are often employed when studying "hidden" or rare populations. These approaches offer alternative strategies and are supported by a growing body of methodological research (Kiesl 2016; Lavallee and Labelle-Blanchet 2013; Medous et al. 2023; Pannuzi and Russo 2014).

5.3.3.10. Longitudinal panel survey/rotating designs

185. When studying changes over time in employment status, for example, it is often more effective to interview the same respondents periodically to capture transitions in employment conditions. This approach is implemented through a panel survey within a longitudinal study design, where participants agree to be contacted repeatedly over time. Panel surveys can be classified into two main types: pure panels and rotating panels, each offering distinct advantages for analysing labour market dynamics.

[1] **Pure Panels:** These involve tracking the same set of respondents throughout the study period. This design allows for the estimation of gross changes, enabling the creation of transition matrices that detail movements between different employment statuses (e.g., employed, unemployed, inactive) (Lynn 2019). Pure panels are commonly used in studies with a defined population and timeframe, such as health or education surveys. However, in continuous surveys, such as labour market surveys conducted by NSOs, maintaining a pure panel is practically impossible. For a survey to remain continuous, respondents would need to participate indefinitely, which is neither feasible nor sustainable. Over time, issues such as respondent fatigue, attrition, and loss of eligibility (e.g., aging out of the labour force) would compromise data quality and representativeness.

[2] **Rotating Panels:** To address these challenges, rotating panels are used in continuous labor market surveys. In this approach, a portion of the sample is systematically replaced over time, following a predetermined rotation scheme aligned with the survey's objectives. Rotating panels strike a balance between maintaining continuity for longitudinal analysis and refreshing the sample to ensure ongoing representativeness and reduce respondent fatigue (Andersson, Andersson, and Lundquist 2011; Kish 1965).

186. Both designs improve the precision of net changes in key indicators, such as employment and unemployment rates, when compared to independent cross-sectional surveys, where each participant is interviewed only once. However, this improvement largely depends on the percentage of overlap, as higher overlap allows for a more accurate measurement of changes over time.

187. Well-established examples of longitudinal labour market surveys include:

- The Continuous Household Survey (ECH) in Uruguay
- The Labour Force Survey (LFS) by Statistics Canada (StatCan)
- The Current Population Survey (CPS) in the United States

188. Rotation patterns vary depending on the reference period used for comparison. For example, the ECH and LFS by Statistics Canada use a six-month rotation system, after which respondents exit the panel. In contrast, the Current Population Survey (CPS) in the United States follows a 4-8-4 rotation pattern, where respondents are interviewed for 4 consecutive months, then rotated out for eight months, and finally re-interviewed for four additional months before leaving the panel entirely.

5.3.3.11. Opt-in panel surveys

189. While longitudinal studies were the original motivation for establishing panels of respondents who agree to be contacted multiple times, it has also become increasingly common to use such groups for cross-sectional studies. When there is a need for rapid responses to a set of important questions, and there is not sufficient time or budget to design a study from scratch, organizations may choose to collaborate with providers managing opt-in panels to include the relevant questions in an upcoming round of interviews.

190. Some opt-in panels are ad-hoc and entirely non-probabilistic. Others are built upon probability samples, where the study organization provides survey responses along with demographic variables describing each respondent and survey weights to align the marginal distributions of respondent characteristics with those of the target population. [[Author Note: Add citations]]

5.3.3.12. Master samples

191. Over time, there has been a growing demand for household surveys in every country. To address this demand efficiently, National Statistical Offices (NSOs) often design a *master sample*, which serves as a comprehensive sampling framework for multiple surveys conducted over an extended period. This approach allows NSOs to draw several simultaneous representative and mutually exclusive samples, which can then be shared with different survey projects, potentially in a rotating fashion.

192. Using a *master sample* avoids the need to draw a separate sample for each survey, streamlining workload planning and making resource allocation efficient. Additionally, because the samples are mutually exclusive, it minimizes the risk of overburdening households with frequent survey requests, thereby reducing respondent fatigue.

193. The master sample may consist of:

- Mutually exclusive sets of primary sampling units (PSUs)
- Mutually exclusive sets of ultimate sampling units (USUs)

194. The NSO managing the master sample will typically share one full sample at a time with survey teams, along with selection probabilities and other relevant metadata about the sampling units. In some cases, survey teams may be asked to share back updated frame-related data fields, such as:

- A fresh count of buildings or occupied dwellings within the sampling unit.
- Updated maps or information about structural or demographic changes over time.

195. A well-constructed master sample provides significant efficiency, consistency, and flexibility in conducting household surveys:

- Efficiency: Reduces costs and time associated with repeatedly building new sampling frames.
- Consistency: Ensures comparability of survey data across different surveys and over time.
- Flexibility: Allows the selection of tailored sub-samples for specific survey objectives while maintaining population representativeness.

196. Master samples are commonly created during intercensal periods to provide a reliable sampling framework between population censuses. They are particularly widespread in low- and middle-income countries, where building a new sampling frame for each survey can be cost-prohibitive and logistically challenging (United Nations 2008a).

197. However, maintaining a master sample requires careful planning and periodic updates to reflect demographic and structural changes, ensuring the sample remains relevant and accurate for ongoing survey needs.

5.3.4. Select a sample design

198. The selection of an appropriate sample design for household and individual surveys conducted by NSOs depends on multiple factors, including:

- The type of sampling frame available (e.g., address-based frames, population registers, area frames, or telephone frames).
- The data collection method (e.g., PAPI, CAPI, CATI, CAWI).
- The analytical objectives of the survey, such as the need for precision at both global and domain-specific levels (e.g., regions, urban/rural areas, sex, age, etc.).
- The survey budget and available resources.
- The frequency and periodicity of the survey (e.g., continuous surveys vs. one-time studies).
- Logistics related to data collection and fieldwork execution.

199. Whenever possible, NSOs should aim to stratify the sample. Stratification enhances both representation and efficiency in survey designs. It ensures that key population groups are adequately represented in the sample and improves statistical precision by reducing variability within strata. Ideally, the strata should align with the survey's domains of interest (e.g., geographic administrative regions or demographic characteristics) to enable reliable estimates across these domains.

200. For many face-to-face household surveys, a multi-stage stratified cluster design is often the most suitable choice. This approach is preferred because up-to-date household frames are often unavailable, and cluster sampling significantly reduces logistical costs associated with fieldwork.

201. For other survey types, NSOs may need to evaluate the available sampling frames, the planned survey mode, and the inferential goals of the survey to propose the most efficient design options. Regardless of the chosen design, it is crucial to balance representation, precision, cost efficiency, and operational feasibility to ensure the success of the survey while meeting its analytical objectives.

5.3.5. Calculate the sample size

202. Calculating the sample size is a fundamental step in designing and implementing a survey, as it directly impacts the quality, reliability, and cost-efficiency of the results. The sample size must be determined by carefully balancing several critical factors:

- Analytical objectives: Ensuring sufficient precision for key indicators to meet the survey's analytical goals.
- Level of disaggregation: Accounting for the domains of analysis (e.g., regions, socio-economic groups, urban/rural areas) to guarantee reliable estimates across subpopulations.
- Budget constraints: Aligning the sample size with the financial resources allocated by the NSO for the survey, including fieldwork, data collection, and analysis costs.
- Expected non-response rates: Anticipating and adjusting for non-response rates to ensure the final achieved sample meets analytical requirements.
- Survey design effect (*DEFF*): Considering the design effect, particularly in stratified or clustered designs, as it affects the required sample size to achieve the desired level of precision.
- Frequency of the survey: Whether the survey is a one-time effort or part of a continuous survey program, as this influences the sample allocation strategy.

203. The key steps involved in calculating the sample size can be outlined as follows:

[1] Identify the main outcome of the survey:
  [a] Determine the key survey outcome for planning purposes.
  [b] Consult with survey stakeholders to understand the required level of precision.
  [c] Define the maximum acceptable margin of error, quantified by the half-width of the two-sided 95% confidence interval.
  [d] If the outcome is a proportion, estimate the anticipated outcome level.

[2] Calculate the effective sample size: Determine the sample size that would achieve the desired precision under a Simple Random Sampling (SRS) design. This is referred to as the *effective sample size*.

[3] Estimate a conservative design effect (*DEFF*): Pay attention to its two components:
  [a] Selection Probability Factor ($DEFF_p$): Reflects different probabilities of selection and adjustments to survey weights.
  [b] Clustering Factor ($DEFF_C$): Reflects the similarity of respondents within clusters.

[4] Adjust for the design effect: Inflate the effective sample size by the *DEFF* to calculate the target number of completed interviews.

[5] Calculate the number of households to contact: Use the expected eligibility rate and expected response rate to inflate the target number of interviews.

204. Comprehensive guidelines and reference materials for survey sample size calculations are available in established resources (International Labour Office 2014; Valliant, Dever, and Kreuter 2013; World Health Organization 2018 Annex B). Furthermore, various specialized software tools—both free and commercially available—are designed to assist with sample size determination for complex survey designs (A. G. Dean, Sullivan, and Soe 2013; DHS Program 2023a; 2023b; NCSS 2024; Thomson, Prier, and Rhoda 2017).

205. Accurate sample size calculation ensures that the survey meets statistical and operational standards while remaining cost-effective and logistically feasible.

5.3.5.1. Effective sample size

206. Because the simple random sample (SRS) is the simplest sample design, it serves as a baseline for comparison with other designs and their properties. Sample size calculations often begin with an equation, software tool, or table that provides an estimate of the required sample size for achieving survey objectives under an SRS design.

207. For a continuous outcome, it is common to invert the calculation for a *Wald-type confidence interval* and estimate the sample size as follows:

$$n_{SRS} \geq \frac{z^2\sigma^2}{d^2} \tag{5-2}$$

where:

- z is the 97.5th percentile of the standard normal distribution (approximately 1.96) for a 95% confidence interval.
- σ represents the anticipated standard deviation of the outcome.
- δ is the desired half-width of the confidence interval, also referred to as the margin of error.

208. For estimating a proportion, the standard deviation can be expressed as $\sigma^2 = p(1-p)$, and the sample size calculation becomes:

$$n_{SRS} \geq \frac{z^2 p(1-p)}{d^2} \tag{5-3}$$

where p is the anticipated proportion of the outcome.

209. When the expected proportion is unknown, a value of $p = 0.5$ is often used because it represents the maximum variance scenario, ensuring a conservative estimate of the required sample size. For example, if the anticipated proportion is 50% (0.5) and the desired confidence interval half-width is no more than 10% (0.10), the calculation becomes:

$$n_{SRS} \geq \frac{1.96^2(0.5)(0.5)}{0.1^2} = 96 \tag{5-4}$$

210. When the expected proportion is smaller than 20% or larger than 80%, the Wald-type confidence interval can become misleading, and alternative methods are recommended for improved statistical properties (N. Dean and Pagano 2015). In such cases, a sample size equation that is based on an asymmetric confidence interval should be considered, following the methodology described in (World Health Organization 2018 Annex B), which incorporates equations from (Fleiss, Levin, and Paik 2003).

211. The resulting value of n_{SRS} represents the *effective sample size*, which is the SRS sample size required to achieve the same level of precision as a more complex design. For complex designs, the effective sample size is typically adjusted (inflated) by an estimated design effect ($DEFF$).

212. While the design effect ($DEFF$) is often greater than 1, leading to an inflation of the sample size, this is not universally true. In certain cases, the design effect can be less than or equal to 1, depending on the sampling design and the homogeneity of the clusters or strata.

5.3.5.2. Design effect

213. The $DEFF$ is a valuable measure for assessing the relative efficiency of different sample designs. Its value varies across different survey outcomes, and it is impossible to know with certainty during the planning stages; it can only be estimated.

214. The $DEFF$ is defined as the ratio of the observed variance of an outcome under a complex survey design to the variance that would result from a simple random sample (SRS) of the same size:

$$DEFF = \frac{Var_{Complex}}{Var_{SRS}} \tag{5-5}$$

215. For planning purposes, $DEFF$ is often approximated as the product of two components:

$$DEFF \approx DEFF_p \times DEFF_c \tag{5-6}$$

where:

- $DEFF_p$: Accounts for the variation in sampling probabilities (e.g., unequal probabilities of selection).
- $DEFF_c$: Reflects a clustering effect that arises when sampling units within clusters are more like each other than units in other clusters.

216. Understanding and accurately estimating the $DEFF$ is crucial for sample size determination and ensuring the precision of survey estimates.

5.3.5.3. *DEFF* due to probability of selection

217. The probability of selection term **($DEFF_p$)** is also known as Kish's design effect (Kish 1965) is typically estimated using data from a recent similar survey, applying the following formula:

$$DEFF_p \approx 1 + CV_w^2 \tag{5-7}$$

where CV_w is the coefficient of variation in survey weights.

218. $DEFF_p$ accounts for variations in survey weights arising from several factors, including:

[1] Differing probabilities of selection.
[2] Differing sampling fractions across strata, including as a special case, oversampling in small populations.
[3] Weighting adjustments:
    - Adjustments for differential nonresponse.
    - Adjustments derived from the use of regression/calibration estimators, such as complete or incomplete post-stratification.

219. These weighting adjustments encompass multiple methods for correcting biases introduced during data collection and estimation, and they should be considered collectively when interpreting the design effect.

220. The Kish design effect formula assumes an optimal allocation strategy, where the sample is distributed proportionally across strata, resulting in equal base weights. However, this assumption does not hold if the variance differs across strata, as unequal variances will impact the precision of estimates.

221. Despite these limitations, the Kish design effect remains a useful approximation, particularly in scenarios where no single variable dominates the survey's analytical objectives – for example, when estimating multiple proportions across different domains. In such cases, the Kish design effect provides a practical baseline for sample size estimation.

222. When varying sampling rates are used across strata—for instance, to maximize efficiency in allocation—the calculated sample size may deviate from what is considered optimal under Kish's assumptions. Therefore, caution is necessary when applying the Kish design effect to inflate sample size estimates, as it may not fully account for the true design requirements of the survey.

5.3.5.4. *DEFF* due to clustering

223. The clustering term, $DEFF_C$, is determined by two parameters:

$$DEFF_C = 1 + (m - 1)\rho \qquad (5\text{-}8)$$

where:

- m is the average number of eligible and completed interviews per cluster.
- ρ *is* the intracluster correlation coefficient (*ICC*), which measures the degree of homogeneity in responses among individuals within the same cluster. The *ICC* is sometimes called the *rate of homogeneity*.

224. The value of *m* is calculated by dividing the total number of completed interviews by the number of responding clusters. In contrast, ρ depends on the outcome variable and may even fluctuate over time. For survey planning purposes, a single conservative ρ value is typically selected based on observed values from past surveys. NSOs often rely on data from their previous survey rounds to estimate an appropriate ρ.

225. It is important to note that $DEFF_C$ increases with both *m* and ρ. In the extreme scenario where each cluster contains only one respondent, the term $(m-1)$ equals zero, resulting in $DEFF_C = 1$. In this case, clustering does not inflate the design effect. On the other hand, excessively increasing m (e.g., once the survey team is on-site, let's maximize data collection in the cluster) can drive up $DEFF_C$, leading to a larger required sample size to maintain precision.

226. The values of the *ICC* are influenced by the degree of outcome variability across clusters, which often corresponds to geographic variability. When an outcome is evenly distributed across clusters (spatial homogeneity), the *ICC* takes its lowest possible value, which depends on *m* and is near zero or slightly negative. Conversely, if all respondents in some of the clusters share the same outcome (e.g., all are mal-nourished) and none of the respondents in other clusters do, the *ICC* reaches its maximum value of 1.

227. Since *ICC* values vary significantly depending on the survey topic, discipline, and country, it is challenging to offer universal benchmarks. NSOs should refer to observed *ICC* values from previous surveys on similar outcomes and consult experienced colleagues to select an appropriate or conservative

value for planning purposes (United Nations 2005 Chapters VI & VII; Korevaar et al. 2021; Dwivedi et al. 2023; Seidenfeld et al. 2023).

5.3.5.5. General comments on *DEFF*

228. In most surveys (including those conducted by NSOs), none of the three key parameters: CV_w (coefficient of variation in survey weights), m (average cluster size), and ρ (intraclass correlation coefficient) are fixed or known until the survey data have been collected. As a result, survey planners often begin with placeholder values, typically derived from previous household or individual surveys or using conservative estimates. These conservative values generally result in a slightly higher $DEFF$ than what may ultimately be observed in the sample. This approach ensures that the precision of key survey outcomes meets or exceeds the target values used during planning.

229. In the context of household and individual surveys, it is often crucial to adopt conservative assumptions because the penalty for imprecision—such as stakeholder indecision or delayed policy implementation—can have far-reaching consequences. Incorporating an appropriate margin of error or buffer in these estimates helps mitigate potential risks associated with imprecise results.

230. While survey planners cannot directly control the value of ρ, they do have some influence over the other two parameters (CV_w and m). For example:

- Homogeneous survey weights (CV_w) tend to have a smaller design effect than widely dispersed weights.
- From a precision perspective, it is generally preferable to have more clusters with fewer respondents per cluster rather than fewer clusters with many respondents per cluster.

231. Sometimes, a proportionate stratified design can yield a design effect smaller than 1, indicating higher precision than a simple random sample. However, in practice, for most household and individual surveys conducted by NSOs, planning for a *DEFF < 1* is not recommended due to the variability and unpredictability of real-world sampling conditions.

232. For additional caveats related to estimating the design effect, refer to (Gambino 2009). This source provides valuable insights and guidelines for handling common estimation challenges.

5.3.5.6. Rules of thumb for DEFF in household and individual surveys conducted by NSOs

233. In household and individual surveys carried out by NSOs, several rules of thumb can guide planners in managing *DEFF* considerations effectively:

- If sampling is conducted directly from individual-level registers or frames, expect lower *DEFF* values.
- When cluster sizes are large, expect higher *DEFF* values.
- Reducing the number of respondents per cluster while increasing the number of clusters can often improve precision.
- Highly variable survey weights will increase *DEFF*. When sampling fractions differ significantly across strata or when differential nonresponse adjustments are applied, survey weights become more variable, inflating *DEFF*.
- Stratified sampling can reduce *DEFF* if strata are well-defined and internally homogeneous.

- Multi-stage sampling designs generally increase *DEFF*. Additional sampling stages introduce more variability, which accumulates at each stage. Efforts should be made to minimize variability in probabilities of selection across stages.

- Oversampling small populations will increase *DEFF*. While oversampling small or specific populations is often necessary for analytical purposes, it introduces variability in weights and increases *DEFF*. Such oversampling should be accounted for in the sample design phase.

- Nonresponse adjustments in weights contribute to *DEFF*. Surveys with high nonresponse rates often require substantial weighting adjustments, increasing variability in survey weights and inflating *DEFF* (*See Chapter 8*).

- *DEFF* depends on survey objectives. If the survey aims to estimate multiple indicators across different domains, conservative *DEFF* assumptions are advisable. Surveys focusing on a single key indicator may allow for more tailored *DEFF* adjustments.

- In cluster sampling, spatial correlation within clusters can inflate *DEFF*. Spatial proximity often increases similarities among respondents, particularly for geographically clustered outcomes. Increasing geographic dispersion of sampled clusters can help reduce spatial correlation effects.

- Conservative assumptions are safer in policy-critical surveys. Surveys providing indicators used for national or international reporting (e.g., poverty rates, unemployment rates) should adopt conservative *DEFF* assumptions to minimize risks of underestimation.

234. These rules of thumb offer general guidance, but it is essential to contextualize *DEFF* assumptions based on the survey design, analytical objectives, and population characteristics. Each survey scenario will require a tailored approach to balance precision, cost-efficiency, and feasibility.

5.3.5.7. Accounting for sampling losses in determining the sample size

235. The final sample size is adjusted to account for the expected eligibility rate and response rate, ensuring that the sample meets the analytical objectives of the survey. To calculate the target number of units to visit, the sample size is further adjusted based on:

- Eligibility Rate (*ER*): The fraction of sampled units (households or individuals) expected to meet the survey eligibility criteria.
- Response Rate (*RR*): The fraction of eligible units expected to consent to and complete the interview.

236. The final sample size can be calculated using the following formula:

$$n_{final} = \frac{n_{SRS} \times DEFF}{RR \times ER} \tag{5-9}$$

237. The eligibility rate can be estimated using demographic data combined with inclusion criteria. For example, if a survey focuses on the immunization status of children aged 12–23 months, demographic data can provide an estimate of how many households survey teams need to visit, on average, to find one child in that age range.

238. The response rate can be estimated from recent similar surveys conducted in the same study areas.

239. It is assumed that the nonresponse mechanism follows a Missing Completely at Random (MCAR) pattern, meaning that nonresponse does not depend on the survey's outcome of interest or auxiliary variables

(Valliant, Dever, and Kreuter 2013). Strategies for managing this assumption through sample size inflation are discussed further in Section 5.3.8. Replacement samples.

5.3.5.8. Sample size – Illustrative example

240. For a survey measuring a proportion under the worst-case scenario (50%) and aiming for a desired precision no wider than ± 3%, the effective sample size can be calculated as follows:

$$n_{SRS} \geq \frac{1.96^2(0.5)(0.5)}{0.03^2} = 1067 \tag{5-10}$$

241. If data from a recent similar survey indicate a CV_w of 0.3, then the design effect due to weighting ($DEFF_p$) can be estimated as:

$$DEFF_p = 1 + 0.3^2 = 1.09 \tag{5-11}$$

242. If recent survey data are unavailable, alternative approaches can be considered:

- If stratification contributes to the dispersion of weights, the CV_w can be derived from the distribution of weights across strata.
- If within-household selection (e.g., selecting one eligible person per household) is the source of variability, the CV_w can be estimated based on the distribution of eligible individuals observed in another survey or census data.

243. If an average cluster size (m) of 10 completed responses per cluster is expected, and assuming an intra-cluster correlation (*ICC*) of 1/6, the design effect due to clustering is:

$$DEFF_C = 1 + \frac{(10-1)}{6} = 2.5 \tag{5-12}$$

The overall design effect (*DEFF*) is calculated as:

$$DEFF = DEFF_p \times DEFF_c = (1.09)(2.5) = 2.725 \tag{5-13}$$

244. Next, if an eligibility rate (*ER*) of 50% is assumed (50% of households are expected to have an eligible respondent), and an anticipated response rate (*RR*) of 85%, the number of households to visit is:

$$n_{final} = \frac{n_{SRS} \times DEFF}{RR \times ER} = \frac{1067 \times 2.725}{0.85 \times 0.50} = 6841.353 \tag{5-14}$$

Rounding up, approximately 6,842 households need to be visited.

245. In a cluster sample with an expected average of 10 completed interviews per cluster, the number of clusters required (m) can be calculated as:

$$m = \frac{n_{final}}{10} = \frac{6842}{10} = 684 \tag{5-15}$$

246. To achieve an average of 10 completed interviews per cluster, it is planned to visit 24 households per cluster. This accounts for the eligibility and response rates:

$$\frac{10}{(0.5\ x\ 0.85)} = 23.5 \rightarrow 24 \tag{5-16}$$

247. These calculations, along with the prospective questionnaire, can be shared with the survey implementation team. They would estimate the budget and time required to conduct the survey under various staffing scenarios.

5.3.5.9. Finite population correction

248. In most surveys, the target sample size represents only a small fraction of the eligible population, making the standard sample size formulas from the previous section appropriate. However, in cases such as surveys conducted in specific small regions, studies targeting small population subgroups, or even when specific stratified domains require separate estimates, the sampling fraction can become significant. In these situations, as the sample size approaches the total population size, the sampling variance decreases, and the desired precision can be achieved with a smaller sample size.

249. This reduction in sample size is achieved by applying the *Finite Population Correction* (*FPC*) *factor*. As a sample approaches a complete census, the variability introduced by sampling decreases because there is less uncertainty about the population's true characteristics.

Adjusted Sample Size Formula with FPC

250. First, calculate the target number of completed interviews as before:

$$n_{eff} = n_{SRS} \times DEFF \tag{5-17}$$

251. If the sampling fraction exceeds 5% of the total eligible population, the sample size can be adjusted using the following formula:

$$n_{adjusted} = \frac{n_{eff}}{1 + \frac{n_{eff}}{N}} \tag{5-18}$$

where N is the population size.

252. For example, if the target number of completed interviews represents 7% of the eligible population, the adjusted target size would be approximately 93% of the original target size.

253. After making this adjustment, calculate the number of households to visit by applying the eligibility rate and response rate:

$$n_{final=} \frac{n_{adjusted}}{RR \times ER} \tag{5-19}$$

254. In the previous example, assume the total eligible population consists of N=9,000 households.

[1] **Calculate the initial target sample size:**

$$n_{eff} = n_{SRS} \times DEFF = 1{,}067 \times 2.725 = 2{,}907.575 \tag{5-20}$$

Rounding up, this suggests a need for approximately 2,908 households.

[2] **Apply the finite population correction:**

$$n_{adjusted} = \frac{2908}{1 + \frac{2908}{9000}} = 2197.85 \tag{5-21}$$

Rounding up, this suggests a need for approximately 2,198 households.

[3] **Calculate the number of households to visit:**

$$n_{final} = \frac{2198}{(0.5) \times (0.85)} = 5171.765 \quad (5\text{-}22)$$

Rounding up to 5,172 households.

[4] **Determine the number of clusters:**
Assuming an average of 10 completed interviews per cluster:

$$m = \frac{5172}{10} = 571.2 \quad (5\text{-}23)$$

Rounding up, this implies a need for approximately 572 clusters.

[5] **Plan the number of households per cluster:**
To achieve an average of 10 completed interviews per cluster, it is necessary to visit 24 households per cluster, accounting for the eligibility and response rates.

255. By reducing the sample size using FPC, the number of clusters decreased from 684 to 582, representing a cost saving of approximately 24% in field operations while maintaining the desired level of precision.

5.3.5.10. Iterating on sample size

256. It is common for a survey steering committee to have unrealistically ambitious goals and hope to obtain very precise estimates for a large number of strata in a very short timeframe with a modest budget. After the first sample size calculation, the committee may learn that what they want would cost far too much or take far too long. Then it is recommended to engage in additional conversation with survey stakeholders to revise the planning parameters and repeat the sample size calculation. (Alternatively, make a successful request for a much larger budget or a much longer timeline!) During the survey planning process, the steering committee for the survey should agree on a set of parameters whose sample size and budget are realistic for a survey whose results will be timely and precise enough to yield useful insights.

5.3.5.11. Pay it forward: Report the observed values of planning parameters

257. When the survey is finished, summarize the observed values of $DEFF$ and m and ρ and CV_w in an annex of the survey report because those values can be very helpful to the team who plans the next survey on this topic in the same country. It may be true that the survey yields a single value for m and for CV_w and yield separate values of ρ and $DEFF$ for each of the main outcome variables. For examples, see (Rhoda et al. 2020 Supplement 2).

5.3.6. Allocate PSUs across strata

258. There are several approaches to allocate primary sampling units (PSUs) across strata. The choice of allocation method depends on the survey objectives, the desired statistical efficiency, the simplicity of implementation, and whether there is prior knowledge about the heterogeneity of the main outcome across strata. In the previous section (5.9.5. Calculate the sample size), the sample size (n) and the total number of PSUs (m) to be selected were determined. This section builds on those results, focusing on how to distribute the m PSUs across the different strata efficiently and in alignment with the survey design objectives.

259. In summary:

[1] **Proportional allocation:** PSUs are assigned to strata in direct proportion to the estimated population size within each stratum. The sampling fraction remains constant across all strata, resulting in similar survey weights. This method is statistically efficient for national-level estimates when the outcome variable does not vary significantly across strata. The goal of proportional allocation is to create a "miniature" representation of the population, ensuring that the weight of each stratum in the sample matches its weight in the population, thus reducing the need for complex weighting adjustments (S. L. Lohr 2021a).

[2] **Equal allocation:** PSUs are equally distributed across strata, regardless of their population size. This method ensures similar precision in estimates for each stratum but is often statistically inefficient for national-level estimates as it produces varying sampling fractions and heterogeneous survey weights. Equal allocation can be particularly appropriate when the primary goal of the survey is to produce region-specific estimates or conduct comparative analyses across strata, as it ensures each region or group receives equal representation in the sample.

[3] **Compromise allocation:** PSUs are allocated unevenly across strata, but not as extremely as in proportional allocation. This approach seeks to balance the benefits of proportional and equal allocation. Methods such as Kish allocation and power allocation aim to optimize the trade-off between efficiency at the national level and precision at the subnational level. These methods are often used when both regional and national estimates are equally important.

[4] **Optimal (Neyman) allocation:** This method uses information on both the variance of the outcome variable and the cost of sampling within each stratum. The goal is to either: minimize overall variance for a fixed total cost or minimize total cost for a given degree of variance. This method determines the optimal sample size for each stratum based on the variance and cost per sampling unit. This approach is particularly effective in surveys where strata exhibit significant differences in variance or cost structures (e.g. urban and rural).

[5] **Square root allocation:** PSUs are allocated based on the square root of the population size in each stratum. This method is often employed when there is limited auxiliary information available and the goal is to achieve approximately equal precision across strata (Särndal, Swensson, and Wretman 1992). It is considered a special case of *power allocation* (Bankier 1988) and is frequently used in scenarios where small domains require sufficient representation to produce reliable estimates.

5.3.6.1. Considerations for disproportionate allocation and oversampling

260. Regardless of the chosen allocation method, disproportionate allocation or oversampling is sometimes necessary, especially in strata where:

- Precise estimates are essential, even if the stratum is small.
- The sampling frame is likely to contain a large number of out-of-scope units (e.g., unoccupied dwellings or individuals who have moved or are deceased).
- There is a higher likelihood of nonresponse, requiring additional sample units to compensate for anticipated attrition.

261. Oversampling is commonly applied to ensure reliable estimates for key subpopulations, such as underrepresented demographic minorities or geographically remote groups. By carefully selecting and applying the appropriate allocation method, survey designers can improve statistical efficiency, precision, and cost-effectiveness, while ensuring that the survey objectives are met across all strata.

5.3.7. Sample coordination and rotation strategies

262. Effective sample coordination and rotation strategies are essential for ensuring the efficiency and sustainability of survey programs conducted by NSOs. These strategies help reduce respondent burden, use resources efficiently, and maintain data quality over time.

263. In the context of master frames (Pettersson 2013) and rotating panels (Cantwell 2008; Lynn 2009; D Steel and McLaren 2008; David Steel and McLaren 2009) sample coordination ensures:

- Minimization of overlapping samples: Avoiding repeated selection of the same households or individuals across multiple surveys.
- Resource optimization: Streamlining fieldwork logistics and reducing operational costs.
- Consistency across surveys: Enhancing comparability of results across survey cycles and domains of interest.

264. For example, an NSO constructs a Master Sample consisting of 2,000 PSUs, stratified by urban and rural domains and other relevant demographic or geographic characteristics. Each PSU contains 20 households, selected under an appropriate sampling design. From this Master Sample, the NSO conducts three major household surveys: a *Labour Force Survey*, which selects 500 PSUs (10,000 households) to measure employment, unemployment, and labour market participation rates; a *Household Income and Expenditure Survey*, which draws 600 PSUs (12,000 households) to collect data on household income sources, expenditure patterns, and living conditions; and a *Health and Nutrition Survey*, which uses 400 PSUs (8,000 households) to evaluate health outcomes, nutritional status, and access to healthcare services. The remaining 500 PSUs (10,000 households) are reserved for new or unplanned surveys that may arise in response to emerging data needs or pilot studies, allowing the NSO to remain flexible and responsive to evolving information demands.

265. The Master Sample ensures minimal overlap between these surveys, as each set of PSUs is carefully selected to avoid repeated inclusion of the same households unless explicitly required for longitudinal analysis. This separation enhances data quality and reduces respondent burden, ensuring households are not repeatedly surveyed within short time intervals.

266. Panel Rotation strategies, on the other hand, allow NSOs to:

- Balance continuity and refreshment: Maintaining longitudinal data while introducing new respondents to preserve representativeness.
- Adapt to analytical objectives: Aligning rotation patterns with the specific goals of each survey, such as measuring short-term or long-term changes.
- Reduce respondent fatigue: Preventing overburdening participants while ensuring sufficient sample overlap for reliable trend analysis.

267. In practice, integrating sample coordination with rotation strategies allows NSOs to achieve:

- Improved precision: Through balanced sample overlap across survey cycles.
- Operational efficiency: With systematic allocation and planned rotation schedules.
- Flexibility: Enabling adjustments to emerging priorities while preserving consistency in survey estimates.

268. For example, in a Labour Force Survey conducted by a NSO, a rotating panel design with quarterly periodicity can be implemented using either a Master Frame or directly from the original sampling frame. The choice between these two approaches depends largely on the statistical infrastructure of the country.

In countries where censuses are conducted every ten years and there is limited capacity to regularly update the sampling frame, a Master Frame becomes a practical and efficient solution. In contexts where the sampling frame can be frequently updated, the rotating panel strategy can be implemented directly from the original sampling frame without requiring an intermediate master sample. Regardless of the approach, the underlying logic of the panel rotation remains the same. For example, using a Master Frame, the NSO constructs a sampling structure consisting of 96,000 households, distributed across 4,800 Primary Sampling Units (PSUs), with 20 households per PSU. This Master Frame is designed to support the rotation strategy sustainably over multiple years, avoiding the exhaustion of available households and ensuring randomness and representativeness in each replacement group.

269. From this Master Frame, the NSO draws an initial sample of 6,000 households, divided into three groups of rotation, each containing 2,000 households (100 PSUs). Each group participates in the survey for three consecutive quarters before being replaced by a new group drawn from the same Master Frame.

270. In the first quarter, Groups 1, 2, and 3 are surveyed simultaneously, covering a total of 6,000 households (300 PSUs). In the second quarter, Group 1 exits the sample and is replaced by Group 4 (another set of 2,000 households drawn from the Master Frame), while Groups 2 and 3 continue. In the third quarter, Group 2 exits and is replaced by Group 5, and in the fourth quarter, Group 3 exits and is replaced by Group 6. This pattern continues consistently, ensuring that each group remains in the sample for exactly three quarters before rotating out permanently.

271. The capacity of the Master Frame (96,000 households) ensures sufficient reserve groups to support this replacement structure over multiple years without repeating households. By maintaining a Master Frame 16 times larger than the quarterly sample size, the NSO avoids the risk of exhausting the available households, thereby preserving the randomness and statistical integrity of the sampling design.

272. Operational efficiency is achieved through predefined replacement groups and geographic clustering of PSUs, simplifying fieldwork logistics and reducing travel costs. Field teams can systematically plan their visits, minimizing interviewer fatigue and ensuring timely data collection.

273. In summary, sample coordination and rotation strategies are useful for maximizing the potential of master frames and rotating panels, ensuring cost-effective, precise, and sustainable survey operations.

5.3.8. Replacement samples

274. In situations where data collection is challenging it can be tempting to select additional sampling units or replace problematic units from the original selection. This practice is generally discouraged, as it can compromise the representativeness of the probability sample. Simply put, outcomes in accessible or cooperative sampling units may differ systematically from those that are inaccessible or uncooperative, introducing non-exchangeable data and potential biases.

275. In most cases, NSOs do not replace entire PSUs. Instead, it is more common to replace ultimate sampling units (e.g., villages or enumeration areas in low-income countries) when access becomes impossible.

276. Replacement of sampling units may legitimately be considered in exceptional circumstances, such as when certain areas become inaccessible due to natural disasters, social unrest, making it impossible to guarantee a safe working environment for survey field teams. In such cases the replacement process should be clearly documented in the Technical Report, including:

- **Mapping:** Maps should indicate which sampling units were dropped and which were added, including the matching criteria used for replacements.
- **Description of limitations:** The Technical Report should include a clear caveat in the study limitations section, explicitly stating that respondents from the dropped sampling units are not represented in the survey results and highlighting this as a potential source of bias.
- **Weighting adjustments:** The report must explain how sampling weights were adjusted for the replacement units, specifying whether their probability of selection calculations accounted for the dropped sampling units (See Chapter 8 for more details).
- **Further methodological guidance:** Detailed procedures for managing replacement scenarios can be found in Chapter 6, which provides additional methodological considerations and best practices.

5.3.8.1. Sample replicates

277. An alternative to direct replacement is the use of *sample replicates***,** a strategy that involves selecting a larger-than-needed random sample based on a 'worst-case scenario' assumption regarding expected eligibility and response rates. In this approach, the sample size is calculated considering the expected eligibility and response rates, ensuring that the final achieved sample meets the survey's analytical objectives even under less-than-ideal conditions.

- **Initial oversampling:** A large sample is randomly selected under a 'worst-case scenario' assumption regarding response rates.
- **Subsampling by replicates:** The full sample is then randomly subdivided into smaller, independent subsamples (referred to as sample replicates). In the case of a two-stage sample, the replicates are usually constructed at the PSU level.
- **Incremental release:** Only the necessary number of replicates is released for data collection to meet the survey's analytical objectives.
- **Preserving randomness:** Because replicates are constructed randomly, withholding any replicate does not compromise the randomness of the overall sample.
- **Weight adjustments:** The weights are adjusted considering only the subsamples (replicates) that were ultimately utilized. The remaining unused replicates are coded as *"unknown eligibility"* to ensure they are appropriately handled during the weighting stage (see Chapter 8).

278. This approach offers several advantages: It ensures that theoretical sample size requirements are met, even in the context of declining response rates and allows NSOs to manage data collection costs effectively while still achieving the precision required for key indicators. For a more detailed explanation of this approach, refer to (Valliant, Dever, and Kreuter 2013).

Application of sample replicates in a two-stage stratified survey design: An illustrative example

279. In a two-stage stratified household survey where the objective is to achieve a sample of at least 1000 completed cases (700 in urban areas and 300 in rural areas) replicates can be implemented to manage data collection efficiently. The design involves selecting 10 households per Primary Sampling Unit (PSU), with an initial oversampling strategy that adds 100 PSUs for the urban stratum and 60 PSUs for the rural stratum. Each PSU serves as one replicate, and once released into the field, it must be fully utilized, ensuring that all selected households are approached. The survey team monitors the number of completed eligible

interviews achieved in each stratum, with particular attention to whether the achieved total is approaching the target (e.g., nearing 700 cases in urban areas). If there are signs that additional replicates might be required, decisions may be made about releasing further PSUs, balancing considerations of time, cost, and field capacity.

280. Fieldwork progress should be closely tracked to avoid releasing unnecessary replicates, as doing so would increase costs without proportional gains in precision. Conversely, delaying the release of replicates can jeopardize the ability to meet survey deadlines. This monitoring is essential, as each replicate represents a predefined sample cluster, and releasing it involves both financial and logistical commitments.

5.3.9. Select the sample of primary sampling units

5.3.9.1. Choose a method of selection

281. Samples are typically drawn from digital sample frames using specialized survey sampling software, which offers advanced capabilities for managing simple and complex sample designs efficiently. These tools facilitate transparency, reproducibility, and adherence to methodological standards in the sampling process.

282. Several widely used software packages include:

[1] **R**: Offers user-written packages for sample selection (Lu and Lohr 2021; Tillé and Matei 2023).

[2] **Stata:** Provides both core commands and user-written routines for sample selection (Jann, Ben 2006; StataCorp 2023).

[3] **SAS:** Features the module PROC SURVEYSELECT, known for its flexibility and advanced options (S. L. Lohr 2021b; SAS Institute Inc. 2015).

283. For replicability and transparency, it is recommended to:

- Save a copy of the sorted sample frame.
- Document and save the syntax used to draw the survey sample.
- Set an explicit random number seed in the syntax before drawing the sample. This ensures that if the selection process is audited or repeated, the exact same sample can be reproduced (Wikipedia contributors 2023).

284. Note: The seed itself does not need to be random. Analysts may reuse the same seed across projects, selecting a favourite integer, or a significant date represented as an integer (e.g., 20250102 to represent January 2, 2025). However, it is recommended to document the chosen seed value before initiating sample selection to prevent bias or manipulation through repeated runs with different seeds.

285. In special situations, when survey-specific software is not available or when public transparency is a priority, it may still be useful to visualize the sample frame in widely accessible tools such as Microsoft Excel or Google Sheets. Performing the selection step-by-step in a public setting can help build stakeholder confidence in the fairness and transparency of the sampling process.

286. Step-by-step guidance for performing simple or systematic random sampling using equal probabilities or probabilities proportional to size (PPS) is widely available in established references, such as (Tillé 2006).

5.3.9.2. Note the probability of selection

287. As detailed in Chapter 8, survey weights depend on the probability of selection at each sampling stage. Therefore, it is essential that:

- The probability of selection is calculated and saved for every sampled unit at each stage of selection.
- These probabilities are accurately incorporated into survey weights to maintain statistical integrity.

5.3.9.3. Preparing the sample file

288. The list of selected sampling units should be saved in a sample file and shared with the teams responsible for:

- Data collection logistics
- Survey weight calculations
- Preparation of analysis datasets

289. The sample file must be accompanied by clear documentation, covering:

- Inferential goal(s)
- Sample design and sample size calculation
- Stratification scheme and PSU allocation
- Sample frame source and description
- Nature of primary sampling units
- Excluded units from the sample frame

290. Each stage of selection should be described and accompanied by the syntax used to draw the sample. If post-stratification adjustments are planned, the source of population total targets should also be documented.

291. Additionally, the sample file dataset should contain:

- Strata, PSUs, and sampling rates for each unit.
- Inclusion probabilities at each sampling stage and the overall probability for each unit.
- Design-based weights with a verification that population size estimates (N) and domain-specific estimates are reasonable.
- If the sample size was increased to account for expected response and eligibility rates, a replicate identifier variable should be included to facilitate data collection.

292. For further details on the processing and handling of this dataset, see Chapter 7).

5.3.10. Elements of the sample file

293. Each sampling unit should be listed with its identification variables, including its strata and any additional relevant details. In the case of area frames, the sample file may include geocoordinates of the PSU boundary vertices or centre, contact information for local officials, and an estimated measure of size (e.g., number of buildings, households, or persons). In the case of person-level frames, the sample file may include postal addresses, email addresses, or phone numbers for sampled individuals. In addition to design and contact variables, the sample file should list the selection probabilities for every stage to facilitate the calculation of design weights.

294. If a stage of selection is conducted by field staff (e.g., selecting from among eligible adult respondents or segmenting an unexpectedly large cluster), it is crucial to provide clear and standardized instructions to the field team. These instructions must outline how to handle such scenarios, and any deviations or adaptations should be clearly documented and reported back to the central team. For example: *There were four eligible adults in the household. A random number table was used to select one of them to participate.*

295. Field staff must not independently decide on a selection method but should strictly follow the protocol provided and report unforeseen challenges or deviations immediately.

5.3.10.1. Recommendations for verifications and system testing

296. Proper verification of the sample file and testing of systems are essential to ensure data quality and operational feasibility.

297. For area sample frames:

- Use GIS software or tools like Google Earth to map PSU boundaries.
- Verify that sampling units do not overlap and are sensible for fieldwork (e.g., not entirely in lakes or uninhabited deserts).
- Assess whether field teams can safely access and navigate the sampled units.

298. For person-level sample frames:

- Cross-reference the sample file with updated lists of individuals who may have moved or died.
- Use external services to obtain updated contact information.
- Verify that data fields used in stratification align with updated information from these services.
- Any discrepancies should be documented and clarified during data collection and weighting processes.

299. The completeness of frame data fields should ideally be evaluated during the frame selection stage. By the time the sample is selected, addressing missing data or incomplete contact information can be challenging. However, documenting the degree of completeness of these fields remains essential, as inconsistencies often emerge during data collection and weight calculation.

300. For further details on how the sample file integrates with data processing workflows see Chapter 7.

5.3.10.2. Generating useful PSU maps

301. For household surveys, generating PSU maps with clearly defined cluster boundaries overlaid on roads, paths, and landmarks is highly beneficial. If the sample frame is georeferenced, this process becomes straightforward. Maps can be generated using:

- Geographic Information System (GIS) shapefiles (ESRI 1998; Wikipedia 2024).
- Keyhole Markup Language (KML) files (Google 2023a; 2023b).

These files are compatible with GIS software (ESRI 2024; QGIS Development Team 2024) and Google Earth (Yu and Gong 2012).

302. Each PSU map can include a text box visible upon selection, listing:

- Project-specific ID labels.
- Metadata relevant to the sampled unit.

303. The spatial layout of these maps facilitates:

- Quick validation of PSU locations.
- Assessment of accessibility for field teams.
- Evaluation of the likelihood of successfully locating respondents.

304. In some cases, organizing a map-a-thon event can be helpful. Volunteers can digitize roads, paths, building footprints, and other landmarks, enhancing the accuracy and usability of project maps (Coetzee et al. 2019; Quill 2018).

5.3.11. Sharing sampling information

305. The full sample file may be shared within the survey team, and much of this information may later be shared with researchers who wish to use the survey data. However, some fields may need to be suppressed or masked to preserve respondent privacy.

306. Downstream data users should have access to a clear and detailed description of the sample design, including any unusual allocations or oversampling strategies that may require special considerations for estimation and analysis. It is particularly helpful to explicitly guide data users on how to describe the sample design in survey estimation software, including:

- Stages of selection: What were the key sampling stages?
- Identification variables: Which variables represent sample strata IDs, cluster IDs, and weights?
- Finite Population Correction (*fpc*): If applicable, which variables represent the population size?
- Weight calibration totals: What was the source of the totals used for weight calibration?

307. If the sample file contains sensitive information, it should be clearly marked for review and evaluated for possible suppression or masking before public distribution.

308. One example of data masking can be found in the USAID Demographic and Health Surveys (ICF 2024). In these surveys:

- PSU geocoordinates are displaced before datasets are released to the public.
- Within the survey team, the true PSU locations are known.
- For public users, the true coordinates are replaced with pseudo-locations that are near the original locations but have been randomly displaced to preserve respondent privacy (Burgert et al. 2013).

309. This approach strikes a balance between data utility and respondent confidentiality, allowing public users to conduct meaningful spatial analyses without compromising privacy.

5.3.12. Later stages in multi-stage selection

5.3.12.1. Selecting households within a cluster

310. The rigor of probabilistic selection can sometimes break down in household surveys at the stage of selecting households within a cluster. The best practice is to generate a fresh cluster-level sample frame that lists all households. This frame should enumerate every household and include a clear map or geocoordinates for locating and revisiting households that are chosen to be part of the sample. The selection process should preferably be carried out by a team or application designed to prevent bias that might arise from interviewer convenience. The probability of selection should be recorded and incorporated into base weight calculations.

311. Some surveys attempt to save time through a "random walk" method, where field teams move through the cluster in a serpentine pattern, skipping households between interview attempts while keeping track of which streets have already been visited. Other approaches select a random starting point and then

visit nearby households, claiming that the randomness of the starting point ensures equal selection probabilities for all households. Less rigorous protocols may begin at a prominent landmark near the centre of the cluster and proceed systematically until a quota of households is visited or a quota of interviews is completed.

312. These methods involve trade-offs. Each approach varies in fieldwork cost and complexity, making simpler methods tempting and efficient. However, different levels of rigor are suitable for different survey goals. For example:

- When outcomes vary spatially within a cluster (e.g., access to services or service usage), protocols focusing only on households near the cluster centre are likely to yield biased estimates (Grais, Rose, and Guthmann 2007).
- If outcomes do not vary spatially, then less rigorous methods might suffice, and small sample sizes per cluster can be more efficient.

313. The United States Centers for Disease Control and Prevention (US CDC) have been developing a tool named *GPSSample* to support rigorous enumeration and sampling within clusters. In some cases, both enumeration and interviews can be completed by the same team within a single day (US CDC 2024).

5.3.12.2. Selecting a respondent within a household

314. When the protocol requires selecting a single respondent from among several eligible individuals in a household, this step should be considered a distinct stage of selection; interviewers should record the number of eligible persons in the household so that weight calculations can accurately incorporate the probability of selection.

315. The method of selecting the respondent will depend on the goals of the survey:

- **Convenient selection** (e.g., interviewing the household head or the first available person) is simple but can introduce bias.
- **Rigorous random selection** reduces bias but adds complexity to the process.

316. Common **methods for random selection** include:

[1] **Most recent birthday method:** The respondent is selected based on whose birthday is closest to the interview date ((Salmon and Nichols 1983). This method is less intrusive and easier to implement compared to others. However, it can introduce bias by over-selecting individuals who are more available or willing to participate. It also assumes that birthdays are uniformly distributed and unrelated to the survey topic, which may not hold true for certain subjects (e.g., spending habits).

[2] **Kish grid method:** Eligible household members are listed in a predetermined order, and one individual is selected using a randomization grid (Kish 1949). This method ensures strict probabilistic selection, where each household member has a non-zero and known probability of selection (Gaziano 2005). However, it may be considered intrusive as it requires collecting detailed information about all household members, which can increase refusal rates (Jabkowski 2017).

[3] **Simple Random Selection (SRS):** One eligible individual is randomly selected from the household, and their probability of selection is recorded.

[4] **Rizzo procedure:** This method balances randomness and intrusiveness (Rizzo, Brick, and Park 2004). It begins by determining the number of eligible individuals in the household. If the person answering the door is among the eligible respondents, a random mechanism decides whether to select them or proceed to another household member using a method like the birthday method. This approach reduces the need to collect detailed information about all household members while maintaining probabilistic selection principles.

See (Binson, Canchola, and Catania 2000; Jabkowski, Cichocki, and Kołczyńska 2024) for comparisons of these methods.

317. In modern survey environments, CAPI can seamlessly program random selection procedures and record selection probabilities, simplifying implementation while maintaining statistical rigor.

5.3.13. Special topics

5.3.13.1. Non-probability samples

318. For NSOs, a probability sample is the recommended approach for most survey applications because it ensures that results are representative and unbiased, and that sampling variability can be quantified. This is essential for producing reliable statistics that meet national and international standards.

319. In contrast, non-probability samples lack these guarantees, as they often suffer from coverage issues and unknown participation probabilities, making it impossible to calculate sampling errors or draw statistically valid inferences to the broader population without relying on assumptions from selection mechanism models (always MAR) or superpopulation models to predict the variable of interest, which, in practice, are impossible to verify or ensure that they hold true for the respondents (Elliott and Valliant 2017; Salvatore 2023; Valliant 2020; Wu 2022).

320. However, NSOs may occasionally employ non-probability samples under specific circumstances, such as:

- **Exploratory research:** To gather preliminary insights before a larger survey effort.
- **Pilot studies:** To test survey instruments or methodologies.
- **Hard-to-reach populations:** When conventional probability sampling is infeasible due to high costs, time constraints, or the characteristics of the target group (Couper, Dever, and Gile 2013; Vehovar, Toepoel, and Steinmetz 2016).

321. In all cases, the limitations of non-probability samples must be explicitly documented in survey reports. NSOs should clearly communicate these limitations to data users, emphasizing the restricted generalizability of findings derived from non-probability samples.

322. Common types of **non-probability sampling in NSO contexts** include:

[1] **Convenience sampling:** Respondents are selected based on availability or ease of access, rather than random selection. While this approach can be quick and cost-effective, it introduces significant bias and compromises representativeness, making it unsuitable for official statistics.

[2] **Quota sampling:** Respondents are selected to meet predefined quotas for certain characteristics (e.g., age, gender, education level). While this method can ensure representation of specific groups, it does not guarantee random selection within those quotas, making inference statistically unreliable.

323. **Best practices for NSOs** when using non-probability samples include:

- Clearly document the rationale for using non-probability sampling.
- Specify the limitations in terms of coverage, inference, and representativeness.
- Avoid using non-probability samples for official statistics or policy-critical estimates.
- Use findings primarily for exploratory analysis or hypothesis generation, not for producing population-level estimates.

324. In summary, for NSOs, non-probability sampling should only be considered when probability sampling is not feasible. Findings derived from such samples must be interpreted with caution and accompanied by transparent documentation of limitations to prevent misinterpretation by data users.

5.3.13.2. Respondent-driven sampling

325. Respondent-driven sampling (RDS) is a specialized method used to study hard-to-survey populations. The process begins with a small sample of initial participants, known as seeds, who then recruit additional members of the target population. Both the seed participants and the new recruits are typically incentivized to encourage participation.

326. Key aspects of RDS include:

- **Recruitment chains:** Newly recruited members can themselves become seeds, creating chains of recruitment that expand the sample.
- **Specialized estimation techniques:** Non-random sampling requires specific statistical methods to adjust for recruitment bias and produce unbiased estimates of population parameters.

327. In theory, these recruitment chains can yield representative and unbiased estimates for networked populations. However, the validity of RDS relies on key assumptions about the structure and connectivity of social networks. When these assumptions are not met (e.g. populations with isolated members) the method's reliability decreases (Gile and Handcock 2010; Heckathorn 1997; Raifman et al. 2022; S. Thompson 2020; Vincent and Thompson 2017).

5.3.13.3. Hard-to-survey populations

328. Some populations are inherently difficult to survey, either due to mobility, invisibility, or sensitivity of their status. These groups may not align well with traditional sampling units in area-based frames. Examples include:

- Nomadic populations or homeless individuals who are not associated with fixed dwellings.
- Individuals excluded from person-level frames due to missing contact information.
- Groups engaged in illicit or illegal activities, who may deny participation or refuse to discuss certain topics (Faugier and Sargeant 1997; Marpsat and Razafindratsima 2010; Tourangeau 2014).

329. Efforts to survey these populations might begin with a probability sample using indirect sampling methods and may subsequently employ respondent-driven sampling (RDS) to expand participation (Gile and Handcock 2010; Heckathorn 1997; Raifman et al. 2022).

330. It is important to clearly distinguish between hard-to-survey populations and indirect sampling methods. While hard-to-survey populations present challenges in terms of frame accessibility and participation, indirect sampling is a technique designed to overcome these challenges through network-based approaches.

5.3.13.4. Balanced sampling

331. Balanced sampling is a method where the sample is selected such that the estimated total or average of auxiliary variables exactly matches the known population parameters while maintaining randomness in selection. Unlike post-stratification, which adjusts weights after data collection, balanced sampling incorporates auxiliary information during the sample selection process. When auxiliary variables are strongly correlated with the survey's key outcome variables, balanced sampling can notably improve statistical precision (Deville and Tillé 2005; Hedayat, Rao, and Stufken 1988; Wright and Stufken 2008; 2011).

5.3.13.5. Adaptive Sampling

332. Adaptive sampling is a dynamic sampling method designed to improve efficiency when studying rare populations or spatially clustered phenomena.

333. Key features of adaptive sampling include:

- **Initial probability sample:** Sampling begins with a traditional random probability sample.
- **Dynamic expansion:** If certain conditions are met during fieldwork—such as finding multiple respondents with the characteristic of interest—adjacent or nearby sampling units are dynamically added to the sample.

334. Adaptive sampling is particularly useful in:

- **Geographically clustered settings:** When respondents with the characteristic of interest are spatially concentrated.
- **Network settings:** When respondents are connected through identifiable social or community networks.

335. Successful implementation requires real-time monitoring of incoming data, precise tracking of selection probabilities, and specialized estimation techniques to avoid introducing bias (S. Thompson 2020; S. K. Thompson 2006; 2012; 2017).

5.4. References

AAPOR. 2023. “Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.” American Association for Public Opinion Research. https://aapor.org/wp-content/uploads/2024/03/Standards-Definitions-10th-edition.pdf.

Andersson, Claes, Karin Andersson, and Peter Lundquist. 2011. “Estimation of Change in a Rotation Panel Design.” In *Proceedings of the 58th World Statistical Congress*. Dublin, Ireland: International Statistcal Institute. https://2011.isiproceedings.org/papers/950903.pdf.

Araujo, Theo, Jef Ausloos, Wouter van Atteveldt, Felicia Loecherbach, Judith Moeller, Jakob Ohme, Damian Trilling, Bob van de Velde, Claes De Vreese, and Kasper Welbers. 2022. “OSD2F: An Open-Source Data Donation Framework.” *Computational Communication Research* 4 (2): 372–87. https://doi.org/10.5117/CCR2022.2.001.ARAU.

Bankier, Michael D. 1988. “Power Allocations: Determining Sample Sizes for Subnational Areas.” *The American Statistician* 42 (3): 174–77. https://doi.org/10.1080/00031305.1988.10475556.

Bethlehem, J.G. 2002. “Weighting Nonresponse Adjustments Based on Auxiliary Information.” In *Survey Nonresponse*. New York: Wiley. https://hdl.handle.net/11245/1.202920.

Biemer, Paul P., Edith Desirée de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker, and Brady T. West, eds. 2017. *Total Survey Error in Practice*. Hoboken, New Jersey: John Wiley & Sons.

Binson, Diane, Jesse A. Canchola, and Joseph A. Catania. 2000. "Random Selection in a National Telephone Survey: A Comparison of the Kish, Next-Birthday, and Last-Birthday Methods." *Journal of Official Statistics* 16 (1): 53–59.

Boeschoten, Laura, Jef Ausloos, Judith Moeller, Theo Araujo, and Daniel L Oberski. 2020. "Digital Trace Data Collection through Data Donation." *arXiv Preprint arXiv:2011.09851*.

Brick, J. M. 2011. "The Future of Survey Sampling." *Public Opinion Quarterly* 75 (5): 872–88. https://doi.org/10.1093/poq/nfr045.

Brick, J. M., P. D. Brick, S. Dipko, S. Presser, C. Tucker, and Y. Yuan. 2007. "Cell Phone Survey Feasibility in The U.S.: Sampling and Calling Cell Numbers Versus Landline Numbers." *Public Opinion Quarterly* 71 (1): 23–39. https://doi.org/10.1093/poq/nfl040.

Burgert, Clara, Josh Colston, Thea Roy, and Blake Zachary. 2013. "Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys." USAID.

Cantwell, Patrick J. 2008. "Rotation Designs and Composite Estimation in Sample Surveys Part 1. Motivating Their Use." U.S. Census Statistical Research Division. https://www.census.gov/content/dam/Census/library/working-papers/2008/adrm/rrs2008-06.pdf.

Coetzee, Serena, Victoria Rautenbach, Cameron Green, Kiev Gama, Nicolene Fourie, Breno Alencar Goncalves, and Nishanth Sastry. 2019. "Using and Improving Mapathon Data through Hackathons." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42:1525–29. https://doi.org/10.5194/isprs-archives-XLII-2-W13-1525-2019.

Couper, Mick P, Jill A Dever, and Krista J Gile. 2013. "Report of the AAPOR Task Force on Non-Probability Sampling." *Retrieved November* 8:2020.

Dean, A.G., K.M. Sullivan, and M.M. Soe. 2013. "Open Source Epidemiologic Statistics for Public Health." April 6, 2013. https://www.openepi.com/Menu/OE_Menu.htm.

Dean, Natalie, and Marcello Pagano. 2015. "Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys." *Journal of Survey Statistics and Methodology* 3 (4): 484–503. https://doi.org/10/gnjws5.

Deville, Jean-Claude, and Yves Tillé. 2005. "Variance Approximation under Balanced Sampling." *Journal of Statistical Planning and Inference* 128 (2): 569–91. https://doi.org/10.1016/j.jspi.2003.11.011.

DHS Program. 2023a. "Calculate Sample Size Using Survey Sampling Online Tools: One Sample." February 16, 2023. https://blog.dhsprogram.com/calculate-sample-size-using-survey-sampling-online-tools-one-sample/.

———. 2023b. "Calculate Sample Size Using Survey Sampling Online Tools: Two Samples." February 16, 2023. https://blog.dhsprogram.com/calculate-sample-size-using-survey-sampling-online-tools-two-samples/.

Dwivedi, Laxmi Kant, Bidhubhusan Mahaptra, Anjali Bansal, Jitendra Gupta, Abhishek Singh, and T.K. Roy. 2023. "Intra-Cluster Correlations in Socio-Demographic Variables and Their Implications: An

Analysis Based on Large-Scale Surveys in India." *SSM - Population Health* 21 (March):101317. https://doi.org/10.1016/j.ssmph.2022.101317.

Elliott, Michael R, and Richard Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32 (2): 249–64.

ESRI. 1998. "ESRI Shapefile Technical Description." ESRI. https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/whitepapers/pdfs/shapefile.pdf.

———. 2024. "ArcGIS." Redlands, CA, USA: Environmental Systems Research Institute. https://www.arcgis.com/index.html.

European Commission. Statistical Office of the European Union. 2015. *ESS Handbook for Quality Reports.* LU: Publications Office. https://data.europa.eu/doi/10.2785/983454.

Faugier, Jean, and Mary Sargeant. 1997. "Sampling Hard to Reach Populations." *Journal of Advanced Nursing* 26 (4): 790–97.

Fleiss, Joseph, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. 3rd ed. New York: Wiley.

Gambino, Jack G. 2009. "Design Effect Caveats." *The American Statistician* 63 (2): 141–46. https://doi.org/10.1198/tast.2009.0028.

Gile, Krista J., and Mark S. Handcock. 2010. "Respondent-Driven Sampling: An Assessment of Current Methodology." *Sociological Methodology* 40 (1): 285–327. https://doi.org/10.1111/j.1467-9531.2010.01223.x.

Google. 2023a. "Keyhole Markup Language | Google for Developers." Keyhole Markup Language. 2023. https://developers.google.com/kml.

———. 2023b. "Keyhole Markup Language - Reference." November 3, 2023. https://developers.google.com/kml/documentation/kmlreference.

Grais, Rebecca F, Angela MC Rose, and Jean-Paul Guthmann. 2007. "Don't Spin the Pen: Two Alternative Methods for Second-Stage Sampling in Urban Cluster Surveys." *Emerging Themes in Epidemiology* 4 (1): 8. https://doi.org/10.1186/1742-7622-4-8.

Groves, Robert M. 2011. *Survey Methodology*. Edited by Floyd J. Fowler, Mick Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2nd ed (Online-Ausg.). Wiley Series in Survey Methodology, v. 561. Somerset: Wiley.

Hansen, Morris H., and Philip M. Hauser. 1945. "Area Sampling - Some Principles of Sample Design." *Public Opinion Quarterly* 9 (2): 183–93. https://doi.org/10.1086/265732.

Heckathorn, Douglas D. 1997. "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44 (2): 174–99. https://doi.org/10.2307/3096941.

Hedayat, AS, CR Rao, and J Stufken. 1988. "Designs in Survey Sampling Avoiding Contiguous Units." In *Handbook of Statistics*, 6:575–83. Elsevier. https://doi.org/10.1016/S0169-7161(88)06026-2.

Himelein, Kristen, Stephanie Eckman, Jonathan Kastelic, Kevin McGee, Michael Wild, Nobuo Yoshida, and Johannes Hoogeveen. 2020. "High Frequency Mobile Phone Surveys of Households to Assess the

Impacts of COVID-19: Guidelines on Sampling Design." World Bank Group. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/742581588695955271/guidelines-on-sampling-design.

Horvitz, Daniel G, and Donovan J Thompson. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47 (260): 663–85.

ICF. 2024. "The DHS Program." The DHS Program. 2024. https://dhsprogram.com/.

ICF International. 2012. *Demographic and Health Survey Sampling and Household Listing Manual*. Calverton, Maryland, USA: Measure DHS. https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

INEGI. 2020. "Guía de Diseño de la Muestra para Encuestas Fase de Diseño de la Norma Técnica del Proceso de Producción de Información Estadística y Geográfica." Instituto Nacional de Estadística y Geografía, Mexico. https://www.inegi.org.mx/contenidos/infraestructura/aseguramiento/doc/guia_de_diseno_de_la_muestra_para_encuestas.pdf.

International Labour Office. 2014. "Sample Size and Margin of Error." International Programme on the Elimination of Child Labour (IPEC). https://www.ilo.org/sites/default/files/2024-09/Child_labour_Sampling_Tools_Tool_01_Sample_Size_and_Margin_of_Error_Guidelines_2014.pdf.

Jabkowski, Piotr, and Piotr Cichocki. 2024. "Survey Response Rates in European Comparative Surveys: A 20-Year Decline Irrespective of Sampling Frames or Survey Modes." *Quality & Quantity*, October. https://doi.org/10.1007/s11135-024-01993-9.

Jabkowski, Piotr, Piotr Cichocki, and Marta Kołczyńska. 2024. "Kish Grid vs Birthday Procedures: Survey Refusal Rates by the Type of within-Household Selection Procedure in the European Social Survey." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 163 (1): 77–87. https://doi.org/10.1177/07591063241258087.

Jabkowski, Piotr, and Marta Kołczyńska. 2020. "Sampling and Fieldwork Practices in Europe: Analysis of Methodological Documentation From 1,537 Surveys in Five Cross-National Projects, 1981-2017." *Methodology* 16 (3): 186–207. https://doi.org/10.5964/meth.2795.

James, Jeffrey. 2021. "Confronting the Scarcity of Digital Skills among the Poor in Developing Countries." *Development Policy Review* 39 (2): 324–39. https://doi.org/10.1111/dpr.12479.

Jann, Ben. 2006. "GSAMPLE: Stata Module to Draw a Random Sample." Statistical Software Components S456716: Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s456716.html.

Kalton, Graham. 2020. *Introduction to Survey Sampling*. 35. Sage Publications.

Kiesl, Hans. 2016. "Indirect Sampling: A Review of Theory and Recent Applications." *AStA Wirtschafts- Und Sozialstatistisches Archiv* 10 (4): 289–303. https://doi.org/10.1007/s11943-016-0183-3.

Kish, Leslie. 1949. "A Procedure for Objective Respondent Selection within the Household." *Journal of the American Statistical Association* 44 (247): 380–87.

———. 1965. *Survey Sampling*. John Wiley & Sons.

Korevaar, Elizabeth, Jessica Kasza, Monica Taljaard, Karla Hemming, Terry Haines, Elizabeth L Turner, Jennifer A Thompson, James P Hughes, and Andrew B Forbes. 2021. “Intra-Cluster Correlations from the CLustered OUtcome Dataset Bank to Inform the Design of Longitudinal Cluster Trials.” *Clinical Trials* 18 (5): 529–40. https://doi.org/10.1177/17407745211020852.

Kreuter, Frauke. 2013. “Improving Surveys with Paradata: Introduction.” In *Improving Surveys with Paradata*, 1–9. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118596869.ch1.

Lavallee, Pierre, and Sebastien Labelle-Blanchet. 2013. “Indirect Sampling Applied to Skewed Populations.” *Survey Methodology* 39 (1): 183–216.

Leeuw, Edith D. de, Joop J. Hox, and Don A. Dillman, eds. 2008. *International Handbook of Survey Methodology*. 1st ed. New York: Routledge. http://joophox.net/papers/SurveyHandbookCRC.pdf.

Lepkowski, James M, and Mick P Couper. 2001. “Nonresponse in the Second Wave of Longitudinal Household Surveys.” In *Survey Nonresponse*, 1st ed., 259–72. New York, NY: Wiley Interscience. https://www.wiley.com/en-us/Survey+Nonresponse-p-9780471396277.

Little, Roderick JA, and Donald B Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

Lohr, Sharon L. 2008. “Coverage and Sampling.” In *International Handbook of Survey Methodology*, 570. New York: Routledge. http://joophox.net/papers/SurveyHandbookCRC.pdf.

Lohr, Sharon L. 2021a. *Sampling: Design and Analysis*. Chapman and Hall/CRC.

———. 2021b. *SAS® Software Companion for Sampling: Design and Analysis*. Chapman and Hall/CRC. https://www.sharonlohr.com/sampling-design-and-analysis-3e.

Lohr, Sharon, and J. N. K Rao. 2006. “Estimation in Multiple-Frame Surveys.” *Journal of the American Statistical Association* 101 (475): 1019–30. https://doi.org/10.1198/016214506000000195.

Lu, Yan, and Sharon L Lohr. 2021. *R Companion for Sampling: Design and Analysis*. Chapman and Hall/CRC. https://www.sharonlohr.com/sampling-design-and-analysis-3e.

Lynn, Peter, ed. 2009. *Methodology of Longitudinal Surveys*. Wiley Series in Survey Methodology. New York, NY: John Wiley & Sons.

———. 2019. “The Advantage and Disadvantage of Implicitly Stratified Sampling.” *Methods, Data, Analyses* 13 (2): 253–66. https://doi.org/10.12758/mda.2018.02.

Marpsat, Maryse, and Nicolas Razafindratsima. 2010. “Survey Methods for Hard-to-Reach Populations: Introduction to the Special Issue.” *Methodological Innovations Online* 5 (2): 3.1-16. https://doi.org/10.4256/mio.2010.0014.

Maslovskaya, Olga, Bella Struminskaya, and Gabriele Durrant. 2022. “The Future of Online Data Collection in Social Surveys: Challenges, Developments and Applications.” *Journal of the Royal Statistical Society Series A: Statistics in Society* 185 (3): 768–72. https://doi.org/10.1111/rssa.12895.

Medous, Estelle, Camelia Goga, Anne Ruiz-Gazen, Jean-François Beaumont, Alain Dessertaine, and Pauline Puech. 2023. “Many-to-One Indirect Sampling with Application to the French Postal Traffic Estimation.” *The Annals of Applied Statistics* 17 (1). https://doi.org/10.1214/22-AOAS1653.

NCSS. 2024. “PASS: Power Analysis and Sample Size Software.” Kaysville, Utah: NCSS, LLC. www.ncss.com/software/pass.

Pannuzi, Nicoletta, and Monica Russo. 2014. “A Methodological Approach Based on Indirect Sampling to Survey the Homeless Population,” no. 1.

Pettersson, Hans. 2013. “Design of Master Sampling Frames and Master Samples for Household Surveys in Developing Countries.”

Peytchev, Andy, Emilia Peytcheva, Johnathan G Conzelmann, Ashley Wilson, and Jennifer Wine. 2020. “Modular Survey Design: Experimental Manipulation of Survey Length and Monetary Incentive Structure.” *Journal of Survey Statistics and Methodology* 8 (2): 370–84.

QGIS Development Team. 2024. “QGIS Geographic Information System.” QGIS Association. https://www.qgis.org.

Quill, Theresa Marguerite. 2018. “Humanitarian Mapping as Library Outreach: A Case for Community-Oriented Mapathons.” *Journal of Web Librarianship* 12 (3): 160–68. https://doi.org/10.1080/19322909.2018.1463585.

Raifman, Sarah, Michelle A. DeVost, Jean C. Digitale, Yea-Hung Chen, and Meghan D. Morris. 2022. “Respondent-Driven Sampling: A Sampling Method for Hard-to-Reach Populations and Beyond.” *Current Epidemiology Reports* 9 (1): 38–47. https://doi.org/10.1007/s40471-022-00287-8.

Rhoda, Dale A., John Ndegwa Wagai, Bo Robert Beshanski-Pedersen, Yusuf Yusafari, Jenny Sequeira, Kyla Hayford, David W. Brown, et al. 2020. “Combining Cluster Surveys to Estimate Vaccination Coverage: Experiences from Nigeria’s Multiple Indicator Cluster Survey / National Immunization Coverage Survey (MICS/NICS), 2016–17.” *Vaccine* 38 (39): 6174–83. https://doi.org/10.1016/j.vaccine.2020.05.058.

Salmon, Charles T., and John Spicer Nichols. 1983. “The Next-Birthday Method of Respondent Selection.” *Public Opinion Quarterly* 47 (2): 270–76. https://doi.org/10.1086/268785.

Salvatore, Camilla. 2023. “Inference with Non-Probability Samples and Survey Data Integration: A Science Mapping Study.” *METRON* 81 (1): 83–107. https://doi.org/10.1007/s40300-023-00243-6.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. Springer Series in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-4378-6.

SAS Institute Inc. 2015. *PROC SURVEYSELECT in the SAS/STAT 14.1 User’s Guide*. Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/onlinedoc/stat/141/surveyselect.pdf.

Seidenfeld, David, Sudhanshu Handa, Thomas De Hoop, and Mitchell Morey. 2023. “Intraclass Correlations Values in International Development: Evidence Across Commonly Studied Domains in Sub-Saharan Africa.” *Evaluation Review* 47 (5): 786–819. https://doi.org/10.1177/0193841X231154714.

Singh, MP, J Gambino, and HJ Mantel. 1994. “Issues and Strategies for Small Area Data.” *Survey Methodology* 20 (1): 3–22.

Skatova, Anya, and James Goulding. 2019. “Psychology of Personal Data Donation.” *PloS One* 14 (11): e0224240. https://doi.org/10.1371/journal.pone.0224240.

Skatova, Anya, Esther Ng, and James Goulding. 2014. “Data Donation: Sharing Personal Data for Public Good.” In *Application of Digital Innovation. London, England: N-Lab*. London, England. https://www.researchgate.net/profile/Anya-Skatova/publication/269101036_Data_Donation_Sharing_Personal_Data_for_Public_Good/links/5480565c0cf25b80dd71101c/Data-Donation-Sharing-Personal-Data-for-Public-Good.pdf.

Skinner, C. J. 1991. “On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys.” *Journal of the American Statistical Association* 86 (415): 779–84. https://doi.org/10.1080/01621459.1991.10475109.

StataCorp. 2023. *Stata Survey Data Reference Manual*. Release 18. College Station, Texas: StataCorp LLC. https://www.stata.com/manuals/svy.pdf.

Steel, D, and C McLaren. 2008. “Design and Analysis of Repeated Surveys.”

Steel, David, and Craig McLaren. 2009. “Design and Analysis of Surveys Repeated over Time.” In *Handbook of Statistics*, 29:289–313. Elsevier. https://doi.org/10.1016/S0169-7161(09)00233-8.

Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald, eds. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. Chichester, West Sussex, U.K: Wiley. https://doi.org/10.1002/9780470688335.

Thompson, Steve. 2020. “New Estimates for Network Sampling.” *arXiv Preprint arXiv:2002.01350*.

Thompson, Steven K. 2006. “Adaptive Web Sampling.” *Biometrics* 62 (4): 1224–34.

———. 2012. *Sampling*. Vol. 755. John Wiley & Sons.

———. 2017. “Adaptive and Network Sampling for Inference and Interventions in Changing Populations.” *Journal of Survey Statistics and Methodology* 5 (1): 1–21. https://doi.org/10.1093/jssam/smw035.

Thomson, Dana R. 2022. *Designing and Implementing Gridded Population Surveys*. Dana Thomson Consulting. gridpopsurvey.com.

Thomson, Dana R., Mary L. Prier, and Dale A. Rhoda. 2017. “WHO Vaccination Coverage Cluster Survey Sample Size Spreadsheet.” Microsoft Excel. https://www.who.int/immunization/monitoring_surveillance/routine/coverage/sample_size_calculator_survey.xlsx?ua=1.

Tillé, Yves. 2006. *Sampling Algorithms*. Springer Series in Statistics Ser. New York, NY: Springer New York.

Tillé, Yves, and Alina Matei. 2023. “Sampling: Survey Sampling.” https://CRAN.R-project.org/package=sampling.

Tourangeau, Roger. 2014. *Hard-to-Survey Populations*. Cambridge University Press.

United Nations. 1984. *Handbook of Household Surveys (Revised Edition)*. Studies in Methods, Series F 31. New York: United Nations.

———. 1986. “Sampling Frames and Sample Designs for Integrated Household Survey Programmes.” DPAJN/INT-84-OI4/5E. NHSCP Technical Study. New York.

———. 2005. *Household Sample Surveys in Developing and Transition Countries*. F 96. New York.

———. 2008a. *Designing Household Survey Samples : Practical Guidelines*. F 98. New York: United Nations.

———. 2008b. “Sampling Frames and Master Samples.” In *Designing Household Survey Samples*, by United Nations, 75–97. Studies in Methods (Ser. F). UN. https://doi.org/10.18356/eba6d79a-en.

United Nations Economic Commission for Europe. 2019. “Generic Statistical Business Process Model (GSBPM).” https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51.

US CDC. 2024. “GPSSample.” 2024. www.gpssample.org.

Valliant, Richard. 2020. “Comparing Alternatives for Estimation from Nonprobability Samples.” *Journal of Survey Statistics and Methodology* 8 (2): 231–63. https://doi.org/10.1093/jssam/smz003.

Valliant, Richard, Jill A Dever, and Frauke Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. 1st ed. Springer.

Vehovar, Vasja, Vera Toepoel, and Stephanie Steinmetz. 2016. “Non-Probability Sampling.” In *The Sage Handbook of Survey Methods*, 1:329–45. 1 Oliver’s Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. https://doi.org/10.4135/9781473957893.

Vincent, Kyle, and Steve Thompson. 2017. “Estimating Population Size with Link-Tracing Sampling.” *Journal of the American Statistical Association* 112 (519): 1286–95.

Welbers, Kasper, Felicia Loecherbach, Zilin Lin, and Damian Trilling. 2024. “Anything You Would like to Share: Evaluating a Data Donation Application in a Survey and Field Study.” *Computational Communication Research* 6 (2).

Wikipedia. 2024. “Shapefile.” Shapefile. 2024. https://en.wikipedia.org/wiki/Shapefile.

Wikipedia contributors. 2023. “Random Seed — Wikipedia, The Free Encyclopedia.” https://en.wikipedia.org/w/index.php?title=Random_seed&oldid=1192099341.

World Health Organization. 2018. *Vaccination Coverage Cluster Surveys: Reference Manual*. March 2021 (WHO/IVB/18.09). License: CC BY-NC-SA 3.0 IGO. Geneva, Switzerland: World Health Organization. https://apps.who.int/iris/handle/10665/272820.

WorldPop. 2025. “WorldPop: Open Spatial Demographic Data and Research.” 2025. https://www.worldpop.org/.

Wright, James H, and John Stufken. 2008. “New Balanced Sampling Plans Excluding Adjacent Units.” *Journal of Statistical Planning and Inference* 138 (11): 3326–35. https://doi.org/10.1016/j.jspi.2006.10.020.

———. 2011. “Variance Approximation under Balanced Sampling Plans Excluding Adjacent Units.” *Journal of Statistical Theory and Practice* 5:147–60. https://doi.org/10.1080/15598608.2011.10412057.

Wu, Changbao. 2022. “Statistical Inference with Non-Probability Survey Samples.” *Survey Methodology* 48 (2): 283–311.

Yu, Le, and Peng Gong. 2012. “Google Earth as a Virtual Globe Tool for Earth Science Applications at the Global Scale: Progress and Perspectives.” *International Journal of Remote Sensing* 33 (12): 3966–86.

CHAPTER 6
DATA COLLECTION

Elizabeth J. Zechmeister (Vanderbilt University)
Elena Cezza (Italian National Institute of Statistics)
Gabriella Fazzi (Italian National Institute of Statistics)
Muhardi Kahar (Statistics Indonesia)
Johannes W. S. Kappelhof (The Netherlands Institute for Social Research/SCP)
Michael Robbins (Princeton University)

CHAPTER 6
DATA COLLECTION

6.1. Overview

1. The overarching goal of Data Collection is to collect data efficiently and effectively, while adhering to local regulations and ethical principles. Efficiency relates to time and budget considerations, while effectiveness includes realizing the intended sample with high quality data. The intended sample refers to who is targeted for interviews from the sampling frame (see Chapter 5). In considering high quality data, two components of the UN Quality Assurance Framework discussed in Chapter 1 are particularly relevant to data collection: accuracy and accessibility. Accuracy refers to minimizing Total Survey Error (TSE) related to representation and measurement quality. Accessibility requires collecting paradata and metadata to document the data collection effort.

2. This chapter is focused on interviewer-administered surveys, primarily face-to-face or telephone. Most surveys conducted by national statistical offices (NSOs) are interviewer-administered (ISWGHS 2024). Parts of this chapter are relevant to self-administered surveys (e.g., web surveys, mail self-completion), but practitioners should consult additional sources for guidance on self-administered surveys.

3. There are three main types of survey data: respondent-provided data (interview), paradata, and metadata. Chapter 1 contains an overview of these types of data. In brief, paradata are supplemental data that document the survey data collection effort (e.g., time stamps, contact attempt outcomes), including deviations to the original design that occur during fieldwork (e.g., a sample substitution) (Kreuter, Couper, & Lyberg 2010). Paradata can be recorded by members of the data collection team (e.g., interviewer observations) and/or can be captured automatically when using electronic tools (e.g., timing data). Metadata are formalized indicators of important characteristics of the survey, such as fieldwork dates and mode (Kreuter 2013; Schenk and Reuß 2024). Metadata are sometimes based on paradata: for example, non-response rates are a type of metadata that are calculated by collecting paradata on successful and unsuccessful contact attempts. Paradata and metadata can constitute key performance indicators (KPIs) that can be monitored and assessed during and after fieldwork.

6.1.1. Guiding principles

4. The UN Fundamental Principles of Official Statistics listed in Chapter 1 converge into guiding principles for data collection around five topics: efficiency, international scientific standards, ethical best practices, confidentiality, and the documentation of practices and results. A core focus of this chapter is scientific best practices to ensure that data collection is efficient (minimizing time and costs, UN Official Statistics Principle 5) and effective – that is, maximizing quality (UN Official Statistics Principles 2 and 9). This section discusses three additional guiding principles related to ethics, confidentiality, and documentation.

5. *Ethics.* Three principles constitute the core of ethical data collection from human subjects: respect for persons, beneficence, and justice (see also UN Official Statistics Principle 2). As detailed in the *Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), these principles have the following prescriptive implications:

- ***Respect for persons*** **-** Inform participants about the nature and purpose of the study so that they can make a considered and voluntary assessment of whether to participate. Take extra care when surveying individuals who have a lower capacity for self-determination (e.g., children, people with cognitive impairments). Be attentive to norms and customs.

- ***Beneficence*** **-** Minimize harms associated with participation, in alignment with notions of reducing risks and respondent burden discussed in Chapters 1 and 4.

- ***Justice*** **-** Broaden opportunities to participate by working to include hard-to-survey individuals or subgroups within the sample, which may require decreasing barriers to entry, respondent burden, or other hurdles to participation (see Chapter 1).

6. *Confidentiality*. Individual-level data collected by NSOs must be kept "strictly confidential" (UN Official Statistics Principle 6). Incorporating confidentiality as a guiding principle for data collection requires planning, monitoring, and assessing efforts to maintain confidentiality during data collection, and adhering to appropriate rules and standards for data security both during and after data collection.

7. *Documentation*. Adhering to strict standards (UN Principle 2) and facilitating correct interpretations of data (UN Principle 3) requires documenting and disclosing information on the study methodology, implementation, and outcomes (protocols, interview data, paradata, and metadata).

8. For more discussion of the principles and guidelines appropriate to survey data collection, see https://wapor.org/about-wapor/code-of-ethics/. Maximizing on guiding principles can have positive consequences for data quality: for example, respecting local norms may increase response rates and reduce bias (while also minimizing risks to the interview team).

6.1.2. Goals

9. The primary objective of data collection is the thorough and standardized implementation of pre-established protocols, including for the questionnaire (see Chapter 4) and the sample design (see Chapter 5). Additional objectives include collecting paradata, adapting to challenges that emerge, staying on time, keeping within budget constraints, minimizing risks, and adhering to principles around ethics and confidentiality. It follows that generating goals for data collection requires special attention to three of the five dimensions of the UN National Quality Assurance Frameworks (see Chapter 1): accuracy (reducing error), timeliness, and coherence (standardization for comparability). It also follows that achieving data collection goals requires training and monitoring a high-quality data collection team.

10. *Implementing the Questionnaire.* The standardized administration of high-quality questionnaires reduces measurement error, decreases item non-response, and increases comparability. Several factors in the data collection process can increase measurement error. For example, interviewers must record responses accurately, which requires training and monitoring for precision in data entry. Section 6.4 discusses factors relevant to training data collection teams in standardized processes for questionnaire administration.

11. *Implementing the Sample Design.* Adhering to the sample design reduces gaps between the sampling frame and the target population (also called coverage error; see Chapter 1). It also reduces the extent to which selected units (e.g., individuals, households) are systematically under- or over-represented (also called nonresponse error; see Chapter 1). This chapter addresses contact and refusal conversion – factors that can affect nonresponse error – in Section 6.6.

12. Errors can be introduced by the interviewer's role in creating the sample. For example, random route / walk approaches involve interviewers in sample construction. Further, some methods of within-household selection are administered by the interviewer; examples include the Kish method and last- or near-birthday methods (see Chapter 5). Interviewer involvement in designing the sample introduces opportunities for both intentional and unintentional deviations in who is recruited into the study (see Battaglia et al. 2008; Bauer 2016; Jabkowski 2017; Kołczyńska et al. 2024). Interviewers (or other members of the data collection team)

also affect the realized sample to the extent they request sample substitutions; when substituted units (e.g., individuals) differ from targets, certain characteristics will be over- or under-represented in ways that bias results (Couper and De Leeuw 2003). One goal is to reduce such the prevalence of such errors, which requires being attentive to vulnerabilities in these methods in training, protocols (e.g., for sample substitutions), and quality control.

13. *Collecting Paradata.* Faithfully executing the study design for data collection requires attention to the collection of paradata and the related production of metadata, including documentation of deviations from the initial protocols for data collection. Achieving this goal necessitates careful planning in advance and monitoring the quality of all data collection processes. Production monitoring and quality control are addressed in Sections 6.7 and 6.8, respectively.

14. *Managing Challenges.* Another goal relates to managing challenges that emerge during data collection. Some challenges can be anticipated, such as the potential for medical, security, or weather-related emergencies in the field; decreases in morale or attrition among the data collection team; and/or lower-than-expected response rates. While challenges that emerge during data collection may be difficult to predict, it is important to establish ways to monitor and respond to potential challenges.

15. *Other Goals.* Other objectives should be derived from guiding principles discussed in the prior section, including goals related to schedule and budget constraints as well as interventions to minimize risks to human subjects and/or data confidentiality.

16. Once all project-specific goals are identified, measures of success – often referred to as Key Performance Indicators (KPIs) – can be developed on the basis of paradata and metadata. For example, specific metrics of success for projects can include target numbers for response rates and numbers of verified and valid interviews (see Chapter 3 and Section 6.7 for more discussion of KPIs).

6.1.3. Historical evolution

17. Modern technology – conventional and smart phones, computers, internet – has revolutionized data collection by introducing new modes and processes (see Chapter 1). Conventionally, most surveys were conducted face-to-face using pen-and-paper interviewing (PAPI) and, to a lesser extent, by mail. The increased prevalence of phones prompted transitions to computer-assisted telephone interviewing. By the 1990s, high income countries began to shift to computer assisted personal interviewing (CAPI), with early adopters in the US, Great Britain, the Netherlands, and Sweden, among others (Banks & Laurie 2000; Martin 1993; Thornberry et al. 1991). While CAPI requires investment in hardware, software, and training, gains in efficiency and quality can quickly compensate for these costs (Caeyers, Chalmers, & De Weerdt 2012). That said, and as discussed in the Emerging Approaches section of this chapter, CAPI does require an assessment of safety risks to bring technology into the field. In countries surveyed by the ISWGHS in 2024, the most common approach is CAPI; some national statistical offices (NSOs) do continue to administer PAPI surveys, either in single-mode and/or in mixed- or multi-mode studies (ISWGHS 2024).

18. Phone surveys, which increased in use over the course of the twentieth century, have likewise deployed new technology to realize computer assisted telephone interviewing (CATI). Computer-assisted interviewing requires specialized training, equipment, software; for example, data collection via CATI can be made more efficient by using private branch exchange (PBX) technology that can auto-dial, transfer calls, and record activity logs. A caveat is that some practitioners find that a pause associated with auto-dialing and call transferring turns respondents away; this outcome that may be more likely in autocratic contexts where a pause could be perceived as a sign that the session is being recorded. Widespread availability of mobile phones has opened opportunities for innovations in CATI surveys around the world.

Where coverage was lacking, early innovations included distributing phones and related equipment (e.g., solar-powered chargers) (Dabalen et al. 2016). As phone ownership increased, attention in LMICs turned to a focus on ways to increase coverage and decrease nonresponse and other errors in the use of phones surveys (for guidance, see Dabalen et al. 2016; Gourlay et al. 2021).

19. As discussed in Chapter 1, the introduction of computer technology into data collection for household surveys reduces errors (e.g., data entry errors), creates efficiencies (e.g., lower interviewer times; Caeyers et al. 2012), and generates paradata that can enhance monitoring and quality control (Kreuter et al. 2010; see also Sections 6.7 and 6.8). Continued evolution and broader access to technology also has generated a proliferation of options for innovative approaches to data collection – including interactive voice response (IVR) interviews; computer-assisted web-interviewing (web); computer-assisted video interviewing (CAVI); interviews utilizing short message services (SMS, or text messaging); and mixed-mode surveys that include methods such as push-to-web (Lynn 2020). The Emerging Approaches section of this chapter discusses some ways that technology is shaping, and will continue to shape, survey data collection.

6.1.4. Roadmap

20. This chapter provides guidance on how to organize and deploy a survey data collection effort in a way that is efficient in the face of omnipresent resource constraints; effective in reducing error; and appropriate in terms of local rules, general ethical principles, and industry standards. The assumption is that both the questionnaire and sample have been designed (Chapters 4 and 5). It is now the task of a data collection team to plan, execute, and monitor the administration of the questionnaire to the intended sample.

21. The remainder of this chapter is organized around the following core subtasks:

- Planning for data collection (Section 6.2)
- Organizing a data collection team (Section 6.3)
- Training (Section 6.4)
- Publicity of survey activity (Section 6.5)
- Data collection implementation (Section 6.6)
- Production monitoring (Section 6.7)
- Quality control (Section 6.8)

22. The chapter concludes with a summary of emerging approaches (Section 6.9) and resources (6.10) for consultation of tools and guidelines for data collection.

6.2. Planning for data collection

23. This section provides a brief overview of these central components of a data collection plan:

- Schedule (Section 6.2.1)
- Mode (Section 6.2.2)
- Sample (Section 6.2.3)
- Questionnaire (Section 6.2.4)
- Paradata (Section 6.2.5)
- Metadata (Section 6.2.6)

6.2.1. Schedule

24. Fieldwork schedules and logistics will vary by project. The schedule should account for holidays, elections, and seasonal / climate factors. Additional considerations include pre-arranging transportation and

security plans for face-to-face fieldwork and scheduling interviewers such that sampled units are contacted on different days and at different times, to minimize bias – specifically, systematic nonresponse error – in the types of individuals who participate in the survey. For more guidance on developing schedules for data collection, see Chapter 3.

6.2.2. Mode

25. The choice of mode(s) of data collection must be made early in the planning process (see Chapter 2). Options include CAPI or PAPI, CATI, web, and paper self-completion (see Chapter 1). New surveys must carefully consider mode options. For regular or repeated surveys, changing mode introduces some risks – e.g., data not being strictly comparable with previous surveys. Nevertheless, it is worth reconsidering data collection modes from time to time to ensure that surveys are as effective and efficient as possible.

26. The optimal mode, or combination of modes, will depend on a wide range of factors. Some modes may be quickly ruled out on the grounds of feasibility, for example low internet penetration rules out a single-mode web survey, and low levels of literacy rules out reliance on self-completion modes. But even when sole use of a particular mode can be ruled out, a mixed-mode design may still be an option – for example if internet penetration is high in one or more large urban areas but low in the rest of the country.

27. Self-completion modes have two key advantages over interviewer-administered modes. One is that, where feasible, they tend to be far less costly. The extent of the cost differential will depend on the methods available to deliver survey invitations and reminders to sample members. For example, if an effective postal service is available, costs can be kept to a minimum. A second advantage of self-completion modes is that for sensitive survey topics, they may elicit more honest responses. The extent to which this is a relevant consideration is survey-specific, depending on the questions to be asked, and context-specific, depending on the nature of social norms.

28. Interviewer-administered surveys, on the other hand, benefit from skills that interviewers bring to the data collection process. Interviewers can respond to sample members' concerns and queries and can persuade them of the benefits of participating resulting in higher response rates than for self-completion surveys. Interviewers can also assist in the interview process by explaining complex or unfamiliar concepts, explaining the response task, and motivating respondents to continue when they are tired of answering questions. This can increase completion rates and boost quality. These strengths of interviewer-administered surveys are somewhat more restricted with CATI rather than CAPI/PAPI, as the interviewer does not benefit from visual communication, such as body language and gestures, and may not pick up on respondent concerns or discomfort. It is also much easier for a sample member to end a phone interview than to dismiss an interviewer who has visited their home in person. CATI brings potential cost savings relative to CAPI/PAPI, but cost benefits can only be fully realized when a good sampling frame is available that includes phone numbers.

29. In many situations, CAPI and PAPI are the only practical options – either because of feasibility or because the advantages they bring greatly outweigh those of other modes. CAPI has some considerable advantages over PAPI and is generally preferred whenever it is feasible. Some of the advantages of CAPI over PAPI are:

- Data entry is not required, eliminating associated costs, time delays, and potential errors
- Routing errors can be avoided as the questionnaire is programmed to follow the correct path through the questionnaire
- Questionnaire structure and routing can be more complex than with PAPI, as all this is handled by the questionnaire program without requiring human intervention

- Data quality can be improved by incorporating "soft checks" in the questionnaire program, so that unlikely responses, or combinations of responses, can be queried immediately with the respondent
- Significantly more paradata can be collected, passively without interviewer effort or room for error, which can facilitate production monitoring and quality control
- Data can be transmitted to the central office as soon as an internet connection is available to the interviewer, which makes for more timely quality control and data processing

30. Use of CAPI, however, requires an initial investment in the necessary hardware (tablets or laptop computers), software (for questionnaire scripting and data transmission), and staff training. Thorough testing of the CAPI system is also necessary before it can be relied on. For this reason, the decision to migrate from PAPI to CAPI is usually made at the institution level rather than being survey-specific. It is somewhat inefficient to run PAPI and CAPI operations simultaneously side-by-side, though many organizations go through a transitionary phase for a period of time, often with different modes being operated in different regions or by different sets of interviewers. CAPI generally becomes an attractive option once the following conditions are met:

- Widespread internet coverage, such that interviewers can transmit completed work to the central office frequently)
- Sufficient computer literacy that it is possible to train a team to use CAPI
- Sufficiently widespread electricity supply to enable to interviewers to recharge devices as required (though portable solar chargers can also be used)
- No significant security risks are posed by interviewers transporting computer equipment

31. Statistical offices and survey organizations relying on PAPI are strongly advised to consider whether and when it may be appropriate to switch to CAPI. The cost-saving, efficiency, and quality benefits could be considerable in the longer run (see, e.g., Caeyers et al. 2012).

6.2.3. Sample

32. While this chapter assumes is already prepared (see Chapter 5), for some approaches certain aspects of the sampling frame are constructed during data collection. This occurs when an interviewer undertakes selection of a sample unit as part of a multi-stage sampling design; examples include designs that include a random route / walk approach and interviewer-administered within-household selection approaches.

33. The following are key steps in developing a plan for realizing the intended sample before data collection begins:

- *Define the sample unit of response and the steps interviewers need to take to reach a potential respondent.* For example, define what constitutes a household and what method will be used for within-household selection.

- *Establish project-specific hardware and software.* For CATI, this may include setting up a PBX system for random digit dialing (RDD). For CATI and CAPI surveys, the sample can be imported into software in advance. For CAPI, recent innovations include placing geofences around the programmed work locations so that deviations from the intended sample areas can be flagged for quality control (see Section 6.8.7).

- *Create project-specific plans for maximizing participation (i.e., the response rate).* This requires attention to groups that might otherwise be under-represented (e.g., individuals in hard-to-reach locations). Plans may include designing advance letters and training interviewers on how to

convert soft refusals into participation (see Sections 6.4 and 6.5). For guidance on hard-to-survey sampling, see Chapter 5; also Tourangeau 2014).

- *Establish protocols for potential deviations (e.g., substitutions) to the intended sample.* For example, if a location or unit cannot be reached, the fieldwork team may propose a similar substitute that is vetted, approved, and documented by the central office. One approach is to have a pre-specified plan for substituting at the household level that can be administered by interviewers without delay; it is recommended, though, that projects require approval for substituting clusters before interviews are conducted in a new location. All deviations from the intended sample must be documented as paradata.

- *Develop options for how to respond to disruptive events.* Extreme weather, social or political unrest, and public health crises are emergency events that can extend fieldwork time, require additional staffing or new training, require substituting inaccessible clusters, require that data collection be paused or postponed, and/or require other interventions by the central office. Fieldwork plans should have an effective system for staying informed and communicating in the event of an emergency. In some cases, a challenge to data collection emerges from multiple survey projects in the field in the same location at the same time; coordinating fieldwork to avoid other projects may reduce respondent fatigue.

- *Develop an approach for recording the outcome of each outreach to a potential respondent.* These contact data are critical to for monitoring the effectiveness and quality of fieldwork; in addition, these data are necessary for the calculation of the survey response rate – the number of units that participate in a survey divided by the total number of eligible units. Final outcomes comprise four primary categories: (1) completed, (2) eligible but not interviewed (non-respondents), (3) unknown eligibility (e.g., a potential respondent is reached but declines prior to screening questions that determine eligibility), and (4) not eligible (e.g., a potential respondent is reached but screening questions determine they are ineligible). Determining what counts as a completed survey requires determining what proportion of questions must be answered to constitute a complete interview (e.g., 80% of all questions); early termination ("breakoff" or "partial interview") rates should be recorded alongside contact attempt paradata.

34. The American Association of Public Opinion Research (AAPOR) provides an exhaustive list of possible outcome codes specified for different sample types as well as different data collection modes (AAPOR 2023). This organization's guidelines are recommended as a frame of reference to increase comparability and standardization.

6.2.4. Questionnaire

35. The questionnaire must be implemented in a standardized way that reduces interviewer and respondent burden. For computer-assisted surveys, after securing the necessary hardware and software (see Chapter 4), the questionnaire needs to be programmed and checked. In addition, when using software for data collection, interview and supervisor accounts need to be created during the planning stage. For face-to-face surveys, planning around the questionnaire includes printing showcards (e.g., with images of Likert scales) and preparing any other materials needed to administer the questionnaire (e.g., for PAPI, the questionnaire itself needs to be printed).

6.2.5. Paradata

36. Paradata is used to monitor the quality of data collection, to reduce sampling, measurement, and other errors (see Sections 6.7 and 6.8). Because paradata can be used in different ways, it is important to identify exactly how paradata will be used.

37. Note that research into interviewer effects may require data collection on the interviewers themselves (e.g., level of experience, age, gender). This requires collecting paradata on interviewers and recording a way to link that information to their interviews. This in turn requires pre-planning around implementation and informed consent for human subjects research.

38. Critically, all potential uses of paradata should be identified in advance and incorporated into the planning stage. For computer-assisted surveys, this planning effort includes programming software to collect the relevant paradata. For PAPI surveys, the paper questionnaire needs to include instructions for the recording of key paradata such as interview date and start and stop times. If employing a production monitoring dashboard (see Section 6.7) to display and assess paradata, that also needs to be prepared in advance.

6.2.6. Metadata

39. Metadata constitute formal and technical documentation describing the data (see Chapter 1). While metadata are compiled after data collection is complete, a thorough data collection plan will incorporate planning for metadata to ensure that information necessary to create metadata is collected. For example, in multi-language projects it is important to record which language each interview took place in; this allows the project to later report a breakdown of the percentages of respondents who took the survey in the different available languages. For details on what metadata ought to be included in technical survey reports, see Chapter 10.

6.3. Organizing a data collection team

40. This section offers information relevant to organizing a data collection team, including:

- Key functions (Section 6.3.1)
- Considerations in team configurations (Section 6.3.2)
- Challenges and solutions: Recruitment and retention (Section 6.3.3)

41. The quality of the data collection team plays a major role in the survey outcome. The primary role of the data collection team centres on engagement with (potential) respondents: recruitment, obtaining consent, interviewing, and (when applicable) recontacting. The data collection team is often referred to as the *fieldwork* team because they are responsible for carrying out (i.e., "fielding") the survey, whether from a central office (e.g., call centre) or in face-to-face interviews in selected locations (i.e., "the field"). It is imperative that the team fully understands the research design, can implement it accurately, and can document and resolve challenges.

42. The specific staffing model will depend on the country, organization, mode, and project. In some cases, team members will be selected from within the NSO/firm; in other cases, data collection team members will be contracted from external firms, and/or the staffing model might require a mix of internal and external individuals.

6.3.1. Key functions

43. This section identifies the functions of a data collection team. For each role, the section offers guidance on requisite qualifications; this guidance can be used to evaluate the quality of existing fieldwork team members and/or to recruit and onboard new team members.

6.3.1.1. Management

44. Managing a survey means being responsible for the overall implementation of the survey (see Chapter 3 on Survey Management). Managing data collection includes developing the fieldwork schedule; leading communication; staffing the data collection team (including making adjustments as needed); recruiting or assigning interviewers; scheduling and/or leading training; overseeing quality control; assessing and mitigating against risks to data protection, fieldwork team members, and respondents; and monitoring production. Managing data collection also requires obtaining necessary approvals; overseeing the budget; securing transportation and/or technology; addressing problems that arise during fieldwork, and leading other tasks outlined in this chapter.

Qualifications

45. Typical qualifications for managing data collection include project oversight experience, managerial skills, and broad knowledge of best practices in survey methodology. Management teams need to be skilled and knowledgeable in finances and regulatory processes. Team personnel must be able to negotiate and navigate contractual obligations' they must be able to ensure that the field team has the resources to carry out their work, that staff are paid on time and in line with their contracts to help maintain team morale, and that all required regulations are known and followed. Ideally, those fulfilling management functions allocate time to stay up to date on advances in survey methodology and have strong networks in the survey research community that can be called upon to help address challenges that arise during planning or implementing fieldwork.

6.3.1.2. Survey sampling

46. During data collection, sampling specialists are involved in monitoring the extent to which the sample is being realized, vetting requests for substitutions, and documenting any deviations from the sample design.

Qualifications

47. Typical qualifications for data collection sampling specialists are specialized training and experience that permit them to design and evaluate samples (see Chapter 5). Data collection sampling specialists might be those same individuals who design the sample or they may be individuals with similar expertise.

6.3.1.3. Data processing

48. During data collection, data processing specialists code, process, and analyze data as it comes back from the field, including monitoring key metrics to ensure that interviews meet expected production and quality control metrics (see Sections 6.7 and 6.8; see also Chapter 7).

Qualifications

49. Typical qualifications for data specialists include specialized training and experience in the statistical analysis of survey data. Strong specialists will be able to the process the data into user-friendly datasets, connect the data to production monitoring dashboards (see section 6.7), create weights, process contact data to generate response rates, and perform various analyses of the survey data (e.g., factor analyses, comparisons of means).

6.3.1.4. Supervision

50. Supervising data collection requires actively overseeing data collection; ensuring adherence to project protocols; and communicating over and resolving issues that arise. For PAPI and CAPI, this includes determining that the interview team has correctly located the selected sampling point/unit; an example of an issue might be an inaccessible location (e.g., gated community), in which case the supervisor needs to submit a request to the operations manager or directly to the sampling specialist for a substitution.

51. The supervising function occupies an intermediary space between management and the interviewing team in two ways. First, the functions fulfilled by supervisors include both managerial and interviewing roles – for example, supervisors may travel with field teams for face-to-face data collection and assist interviewers in fulfilling their assignments). Second, supervisors act as a communications link that connects management and interviewers – for example, supervisors may be the first to be alerted to a problem, and may need to report on that issue to management.

52. Supervisors directly and/or indirectly observe interviews. For face-to-face surveys, some projects assign four interviewers to a supervisor, but numbers vary. In face-to-face surveys, supervisors are often responsible for communicating with local authorities and overseeing the team's safe and efficient transport. When using CAPI for production monitoring and quality control (see Sections 6.7 and 6.8), fewer supervisors may be needed for direct (in-person) observation in the field; rather, staffing needs to ensure a roster of individuals who can monitor the electronic data secured by the CAPI approach. When problems or risks (e.g., to data quality or confidentiality and/or to the safety of the team) emerge in the data collection process, it is the function of supervisors to resolve these challenges and, if necessary, bring them to the attention of management.

Qualifications

53. Qualifications for supervisors vary, but often effective supervisors are themselves high quality interviewers. This is useful because they may help recruit and/or (re)train interviewers; coach struggling interviewers while data collection is in progress; and (when relevant) perform interviews themselves. Supervisors should have good problem-solving and good communications skills.

54. The ideal supervisor can act as both a coach and an authority figure. To maintain interviewers' morale and effectively coach interviewers, the ability to coach high quality performance is essential. High quality coaches can improve the performance by role modelling, encouraging interviewers, providing feedback for improvement, and helping interviewers learn throughout the process. However, when needed, supervisors must be willing to act as an authority: if the supervisor sees evidence of particularly low data quality or outright fabrication, the supervisor must act urgently and decisively.

6.3.1.5. Training

55. The function of trainers is to design, carry out, and assess interview trainings (see Section 6.4 on types of trainings, such as general vs. project-specific). Tasks assigned to trainers may include creating or revising manuals, designing interactive exercises, and conducting post-training assessments. While most trainings are centered on the role of interviewers, it is important for the staffing plan to anticipate any needed trainings of those performing other functions (e.g., production monitoring and quality control).

Qualifications

56. Qualifications for trainers include in-depth knowledge of general and project-specific survey protocols; appropriate language skills; and training or experience effectively teaching adult learners. Trainers should be aware of the social (cultural) context of the survey and be knowledgeable about ethical

principles for human subjects research. As noted above, this may be a role that is assigned to a manager or supervisor. Training is discussed in more detail in Section 6.4.

6.3.1.6. Interviewing

57. Interviewers serve as the primary point of contact with target respondents and fulfill two essential functions: persuading target respondents to participate (thereby reducing sampling and non-response error) and clearly and accurately administering the questionnaire while faithfully recording respondents' answers (therefore minimizing measurement errors and processing issues). Interviewers may also play a role in sample construction, for example in random route / walk designs and/or in designs that require interviewers to perform within-household selection of respondents. Recognizing the interviewer's impact on the nature and outcome of the interview is an essential first step in staffing a project (see Randall et al. 2013).

58. Generally, strong interviewers share several traits. They are effective communicators, who can read questions accurately and clearly. They have strong interpersonal skills and ability to navigate difficult or stressful situations. For example, they need to be able to manage cases where respondents have adverse reactions to being asked to participate or to some of the specific questions in the survey. Interviewers who are more confident, positive, and/or extroverted can potentially achieve higher response rates (West and Blom 2017). It is worth noting that there is typically variability in the strengths that interviewers have across different tasks: e.g., some might be better at contacting and recruiting, while others excel at faithfully administering the questionnaire as it is written (Morton-Williams 1993; Olson and Bilgen 2011).

59. Interviewers need to be adept at using the materials and, where relevant, technology required to carry out their role. For certain surveys, they will need to transcribe open responses efficiently and effectively and/or they will need to be adept at using mobile devices or other technology.

60. For multi-language projects, it is important to select and assign interviewers based on language capabilities. Language mismatches can lead to failed or low-quality interviews (Nguyen, 2025). In some cases where many languages are spoken, such as India or some sub-Saharan Africa countries, the fieldwork organization may have to do their best to match respondents with interviewers who speak their language based on geography or find other workarounds.

6.3.1.7. Public engagement

61. Another important function for a data collection team is to publicize, answer questions, and/or take feedback from potential respondents and/or the public (see Section 6.5). Survey projects will vary in their specific approaches. Some will use commercial options to advertise a study to the public and provide a webpage, QR code, telephone number, or other means of contact. Most projects will provide contact information to potential and actual participants.

Qualifications

62. Qualifications will vary based on the nature of the publicity effort (see Section 6.5). For example, those staffing a call centre or responding to web contacts should be able to address questions about the survey, whether from the public, media, officials, and/or (potential) participants and have a system for reporting any complaints that arise from the fieldwork process.

6.3.2. Considerations in team configurations

63. Surveys are conducted by diverse types of organizations, including NSOs, other government ministries, private firms, international/regional organizations, among others. There is significant variation, both across and within these organizational types, in how data collection teams are organized. Below are five key dimensions, along with some trade-offs, in how data collection teams are organized:

- *Hourly vs. completed interview payment schedules:* Scheduling compensation on the basis of completed interviews may increase efficiency (speed), at the cost of quality (accuracy). (Kiecker & Nelson 1996; Menold et al. 2018; Winker et al. 2015). If an hourly rate is not possible, a payment schedule should account for all tasks required of data collection team members (e.g., filling out contact forms); a further option is to provide bonus payments to reward high quality work.

- *Local vs. non-local interviewers:* Especially for CAPI or PAPI surveys, it may be more efficient to recruit and train interviewers in a central location and then send them to various localities; further, using non-local interviewers who are not embedded in local social networks may increase safeguards on respondent confidentiality. Yet, using local interviewers may decrease travel costs and, due to shared customs, dialects, or other factors, may have better success in securing cooperation from respondents (SRC 2011; Weinreb et al. 2018).

- *Experienced vs. inexperienced interviewers:* More experienced interviewers may be more efficient – that is, they may have lower interview times (Olson and Peytchev 2007). They may also be better at establishing good rapport, boosting response rates (Groves and Couper 1998). Yet, they may simultaneously increase acquiescence (respondent agreement with questions) in ways that affect data quality (Olson and Bilgen 2011) and they may be more prone to deviate intentionally from protocols (Doušak et al. 2024). In contrast, less experienced interviewers may be less efficient and make more unintentional errors (Doušak et al. 2024; Olson and Bilgen 2011; see also Wuyts, 2022). Knowing that level of experience can affect interviewer performance across different tasks has implications for approaches to training (e.g., working with less experienced interviewers on rapport-building) and quality control (e.g., checking to ensure more experienced interviewers adhere to protocols).

- *Traits that boost recruitment vs. traits that boost data quality:* Traits that make interviewers more successful in recruiting respondents (e.g., agreeable personality) may negatively affect data quality (via intentional deviations from protocols to keep up rapport with respondents or via boosting respondents' own agreeableness (leading to higher response acquiescence), and vice versa (Morton-Williams 1993; Olson and Bilgen 2011).

- *Diversity vs. homogeneity:* Configuring a diverse team is consistent with goals around inclusivity (see Chapter 1) and diffuse the potential for any one type of interviewer effect to introduce a large bias, yet a particular project may be more efficient and effective if interviewers have a restricted set of traits – for example, if the target population shares a set of traits that may affect cooperation, then it may be desirable for the interviewers reflect those traits (see, e.g., Kappelhof and De Leeuw, 2017). Further, some projects will have very specific requirements stemming from the nature of the survey – e.g., individuals with specialized knowledge relevant to certain topics (e.g., agriculture, enterprise) or individuals with particular traits (e.g., all female). Decisions over how whether to consider the background traits of interviewers should made with an awareness of the potential for interviewer effects – that is, for recruitment outcomes or question responses to be correlated with specific interviewers or their traits.[1]

[1] Interviewer gender, age, level of experience, race, and religion can shape responses, though the specific nature of these outcomes varies by study (for an inventory of many of these studies, see the detailed appendix provided by West and Blom 2017; on religion, see Blaydes & Gillum 2013; Mneimneh et al. 2020). Quasi- or complete random

64. Project-specific constraints affect choices around the above trade-offs. For example, in multi-language and/or geographically-challenging contexts, using local interview teams is often comparatively more practical. Being aware of trade-offs can inform decisions about how to configure teams and allocate resources. For example, when using local interviewers, a train-the-trainer approach may reduce costs (see Section 6.4). Or, it may be that an internal roster of experienced interviewers already exists; in this case, it may be prudent to devote fewer resources to training, and more resources to supervision and quality control (see Section 6.8). Resource-permitting, another approach to addressing trade-offs is to conduct pilot tests (e.g., test whether interviewers with a particular characteristic are more efficient at recruitment, while paying attention to any potential changes in data quality).

65. [[Placeholder for PENDING: NSO CASE EXAMPLES/STUDIES FOCUSED ON CHALLENGE AND SOLUTIONS]]

6.3.3. Challenges and solutions: Recruitment and retention

66. Some projects require hiring new data collection team members and/or contracting individuals or firms. Whether working with an existing roster of team members or recruiting new personnel, retention of personnel is an important challenge. Team turnover creates inefficiencies and other challenges for data collection. Some considerations for maintaining a stable roster are ensuring that compensation is fair such that it is motivating and creating an organizational culture that supports and encourages team members. One way to transfer knowledge and motivate personnel is to host (de)briefing sessions in which team members exchange best practices.

67. Below are additional examples of how organizations have addressed challenges related to recruitment and/or retention. [[PENDING: NSO CASE EXAMPLES/STUDIES FOCUSED ON CHALLENGE AND SOLUTIONS]]

6.4. Training

68. The critical role played by training in boosting efficiency (i.e., rate of productivity) and effectiveness (i.e., data quality) has been long-recognized and researched (United Nations 1984; Billiet and Loosveldt 1988; Iarossi 2006). For example, well-trained interviewers achieve higher response rates in general (Daikeler and Bosnjak 2020) and in particular on sensitive topics; they are more efficient in providing definitions (when permitted) for difficult questions and in utilizing probing techniques (Billiet and Loosveldt 1988). While a core emphasis is on the training of interviewers, this section also provides guidance on training other key roles, including supervisors, trainers, and data processing specialists.

69. Training approaches will vary by project, with decisions influenced by mode, size and scope, language(s), and other considerations. This section discusses options that teams can use to develop an effective approach for a given survey. Specifically, this section addresses:

- Training approaches (Section 6.4.1)
- Number and type of trainees (Section 6.4.2)
- Training techniques (Section 6.4.3)
- Training logistics (Section 6.4.4)
- Training materials (Section 6.4.5)

assignment of interviewers to target respondents may diffuse interviewer effects and, under certain designs, permit estimations of interviewer effects (see, Kappelhof and De Leeuw, 2017).

- Training content (Section 6.4.6)

6.4.1. Training approaches

70. The optimal training approach for a given study will vary by project characteristics, including mode, size, language(s), and other considerations. Regardless, two considerations are essential. First, logistics must be planned in advance. For example, most trainings require adequately sized classrooms and well-prepared training materials such as paper printouts, e-slides, and/or equipment such as computers and tablets. Second, training sessions within projects must be standardized to minimize variability in outcomes (see discussion in Kutka et al. (2023, web).

71. Training approaches can be categorized into two models: centralized and decentralized. It is common for decentralized trainings to use a train-the-trainer (TTT) method: central representatives train a set of trainers, who subsequently train subsets of supervisors and interviewers with similar characteristics, whether by location, language, or other factors.

72. A *centralized approach* typically involves bringing all trainees to a single location. It is ideal for smaller teams who communicate in the same language. A key benefit to a centralized approach is standardization, because all interviewers are trained by the same trainer(s). In addition, centralized trainings can facilitate direct interaction between survey designers and the data collection team. This interaction can promote consistency in information conveyed and enhance morale and motivation (Grassi and Romano, 2020).

73. A key consideration in selecting the centralized approach is whether it is practical to bring all trainees to a single location. Centralized approaches work well for CATI teams that work out of a single facility (call centre) or CAPI teams in smaller countries. It can also be particularly effective when the interviewing team consists of individuals from a single organization; when the pool of interviewers reports directly to the study director, the duration and location of trainings can be decided internally and then updated as needed on a project-specific basis. Conversely, centralized approaches can be expensive or impractical if interviewers are distributed across different geographic areas, across organizations (such as local networks), or speak different languages.

74. A *decentralized approach* to training brings groups of trainees to different locations. For large projects, multi-lingual projects, and projects for which interviewers come from different geographic locations, it can be cost-saving to use a decentralized approach. A key benefit of a decentralized approach is that, by training smaller groups, the sessions may permit greater interaction between trainee and trainer. While decentralized trainings can be carried out by the same trainer(s) traveling to different sites, it is common for decentralized approaches to use a TTT method.

75. A key consideration in selecting this approach is that, because decentralized approaches mean different individuals are engaged in each training, the quality of the trainers' instruction becomes crucial to ensure uniformity and consistency in the messaging, so that, ultimately, the data collection processes are standardized. Especially when decentralized trainers (e.g., TTTs) are less experienced, it is important to devise training materials and training assessments that ensure all trainees receive the same level of high quality, standardized training.

76. Training may be conducted by the NSO or by an external firm. Two considerations are important when data collection is carried out by an external firm. First, all decisions regarding training logistics must be made during the definition phase of the call for tender, including details on the type, quantity, and other expectations (Daikeler and Bosnjak, 2020, p. 62). Second, if outsourcing training to an external firm, it is important to careful evaluate that firm’s capacity before and during the training. External firms may be

tempted to abbreviate training methods as a cost-savings technique and, instead, rely on interviewers' prior experience. Individuals who are not trained internally may lack knowledge or motivation to adhere to the high-quality standards that derive from the UN Fundamental Principles of Official Statistics that are listed in Chapter 1 and related best practices identified in this manual.

77. The primary training approach – centralized or decentralized – may be complemented by additional trainings during data collection. All projects should plan for how new team members will be trained. It is not uncommon for a project to need new interviewers due to voluntary attrition (as discussed in the prior section, staff retention is a challenge), dismissals for quality control reasons (see Section 6.8), and/or pacing issues or schedule constraints that make it necessary to increase interviewers or other data collection team members to expedite data collection. Organizing extra training sessions during ongoing data collection is challenging. One approach is to have new interviewers independently read materials and view prerecorded training sessions asynchronously (on their own time), and then meet with supervisors face-to-face (synchronously) for sessions that assess what they have learned and focus on practicing interviewing skills.

6.4.2. Number and type of trainees

78. It is advisable to train a surplus of individuals for roles like interviewers and supervisors. This facilitates competency-based selection of the actual data collection team, compensates for attrition, and provides a trained reserve for quick substitution as needed. Kutka et al. (2023) recommend training at least 20% more fieldworkers than needed if working with known, reliable teams, though some recommend even more (see Schnell, 2012; Stiegler and Biedinger, 2016). For new trainees, or high dropout scenarios, it may be useful to increase the surplus up to 40% above the projected team size (Kutka et al., 2023). Training an oversupply increases both organizational burden and budget and, thus, decisions will vary by project. Data from similar prior studies, including on interviewer attrition rates, can help determine the optimal number of trainees.

79. Levels of expertise and experience can vary across types of trainees. In some cases (e.g., rolling and repeated surveys), it is ideal to conduct two types of trainings: one for experienced team members that covers specific changes in the particular survey and one for new team members that covers both general techniques and the specifics of the survey (see discussion in Balì et al., 2023a; see also Deikler et al. 2017). Yet, projects should not lose track of the fact that experienced interviewers need to be reminded of project protocols. Further, facilitating dialogue between experienced and new interviewers can build team culture and allow those with experience to share valuable insights (Cohen and Lotan, 2014).

6.4.3. Training techniques

80. Three choices in training techniques are the following: interactive vs. static lecture, in-person vs. online, and synchronous vs. asynchronous. Keep in mind that the average adult’s capacity to pay attention is limited. Thus, it advisable to schedule breaks at least every two hours (Kutka et al. 2023) and to alternate between different types of content and activities (Knowles et al. 1984).

81. *Interactive* approaches hold many benefits over *static lectures*. and particularly so when material is presented in a static fashion. Interactive experiences boost memory retention and internalization more than simple listening; further, ample practice opportunities enhance capacity to implement protocols in the field (Kutka et al., 2023). A meta-analysis by Daikeler and Bosnjak (2023) finds that, when interviewer training includes practice and feedback sessions, response rates are on average 13% higher.

82. Interactive activities can be built into trainings in different ways. Incorporating short quizzes during lectures promotes engagement, discussion, and identifies areas that need reinforcement (Ketge, 2024). Trainers can leverage technology for these types of activities. For example, quick polls can be launched via

apps or web platforms to keep trainees attentive. To encourage interaction among trainees, individuals can take turns reading materials, such as questionnaire modules, to the group. Kutka et al. (2023) provide the following tips to boost trainee engagement: "ask frequent questions, have trainees read questions and manuals aloud, encourage them to correct each other. Avoid always choosing the most engaged trainees and try to involve everyone by using a random name picker."

83. Importantly, training sessions should include mock interviews in which participants take on roles as interviewees and/or interviewers in realistic scenarios – e.g., converting a hesitant potential respondent or interviewing an impatient respondent. These activities build confidence and comfort, with the goal of improving recruitment and interviewing skills. In addition, a field preparation activity (rehearsal), with real-life interviews in the field, can be held at the end of training; in these cases, all arrangements should be made in advance so that time in the field is spent on training (Kutka et al., 2023). In some cases, trainees should practice contacting potential respondents, in particular if they are involved in the sample selection.

84. *In-person* learning has the benefit of direct interaction with trainers and peers, and real-time feedback, but if the network of trainees is diffuse and/or large, it can require more organization and a higher budget for transportation, site rental, and/or other costs. *Online training* (also referred to as distance or remote training) developed as a practice in the 2010s and accelerated in 2020 due to the COVID pandemic (see Balì et al., 2023a). In some contexts (in case of war, or health emergency), remote learning is the only available option. In the experience of Palestinian Central Bureau of Statistics, the training of Gaza interviewers has often been carried out through video-conference sessions and recorded materials. While online training can reduce costs, trainees may be less engaged and less satisfied (compared to face-to-face; see Balì et al., 2023b) and some learners will find it more difficult to ask clarifying questions in remote versus face-to-face-classes. Other research has found that in-person vs. online trainings can yield similar outcomes, yet that is conditional on the careful design of each approach (Balì et al., 2023b).

85. Remote training sessions should be kept short and include ample discussion elements and strategies to reduce the distance, using tools that allow greater interpersonal connectivity, such as icebreakers (Balì et al., 2023b). At least once an hour, trainers should pause to check in, such as by asking a question that everyone needs to answer; further, trainers can use group discussions on video-chats to keep participants active and engaged (Roll-Hansen and Rustenbach, 2021). Given that it is comparatively hard to monitor attentiveness, distant learning approach should prioritize assessments. For example, each session might be followed by a supervisor reaching out to each interviewer for a brief one-on-one voice over internet protocol (VoIP) or regular phone to ensure they were attentive during training and gained a clear understanding of the task.

86. Learning is *synchronous* when material is covered by the instructor and trainees at the same time, whether online or in-person. Synchronous training often requires less preparation time. Learning is *asynchronous* when material is covered by trainees before or after engagement with the instructor. Simple forms of asynchronous learning include providing trainees with material to review in advance so that the synchronous meeting can focus on discussing the material rather than presenting it for the first time (also known as "flipping the classroom"). Online approaches offer opportunities for entirely (100%) asynchronous approaches. To be effective, such asynchronous training requires the creation of comprehensive materials and video recordings, along with additional time for video and audio editing to ensure quality.

87. In choosing training techniques, it is possible to combine different elements. A blended learning approach may include asking learners to read a text, complete an assignment, take a quiz, watch a video, or attend a lecture or workshop. Blended learning can leverage the flexibility of online learning - regarding

time, location, and cost - along with the benefits of face-to-face instruction and some research suggests this is a particularly effective way to administer trainings (Daikeler and Bosnjak, 2023).

6.4.4. Training logistics

88. The appropriate number and size of training sessions varies depending on factors such as the nature of the interviewer team (e.g., size and experience), sample size, geographical dispersion, survey duration, complexity of the questionnaire, language(s), and data collection methods. Smaller groups will be more conducive to interactive learning approaches. What constitutes a small group can depend on the skill and capacity of the trainer(s) to elicit engagement. When deciding to hold multiple training sessions, it is critical to prioritize standardization by using the same materials, activities, and assessments.

89. When groups of trainees are large in number (e.g., more than 50), some options include:

[1] *Plenary + break-outs:* A plenary session for all trainees for less size-sensitive instruction modules and break-out sessions that divide trainees into smaller groups for practical exercises and size-sensitive activities

[2] *Parallel training in one location:* Trainees are split into multiple classes and trained simultaneously at the same place with a dedicated trainer for each class and additional trainers rotating between groups to maintain a high standard of content consistency

[3] *Parallel locations training:* Trainees are split into multiple classes located in different places, requiring multiple trainers (consistent with the train-the-trainer [TTT] approach discussed in subsection 6.4.6)

[4] *Consecutive training:* Trainees are split into multiple classes and trained consecutively in the same or different locations with the same team of trainers delivering a standardized experience for each group

90. The choice of training schedule is independent of decisions regarding training techniques (lecture vs. interactive, online vs. in-person, synchronous vs. asynchronous), yet the effectiveness of the sessions can be boosted by careful attention to training techniques. For example, if selecting to conduct a *plenary + breakouts* training schedule, one could have individuals review material asynchronously, prior to attending the plenary, and then incorporate an interactive discussion or quiz on that material during the plenary session.

91. Training sessions for projects should be held close to the data collection period—ideally one to three days before fieldwork begins—to ensure that skills and techniques remain fresh and easily applicable (Stiegler and Biedinger, 2016). Because adhering to this recommendation can be challenging with a consecutive training approach, it may be important to offer refresher material immediately prior to the start of data collection in those or other cases in which training sessions are not held immediately prior to the start of data collection.

92. The optimal duration of training varies based on several factors, including the experience level of interviewers, the roles involved, and the complexity and length of the questionnaire. There is no universal standard. For new or complex projects, some recommendations include five-day training sessions (Iarossi, 2006) while others recommend a longer training of at least 15 days for a 2.5-hour questionnaire (Kutka et al., 2023); in a project on displaced persons led by the University of Michigan, the training duration was four weeks. The daily hours dedicated to training also differ depending on the format; typically, the number

of hours per day should be lower for online classes versus in-person sessions (though, even in these latter cases, breaks are necessary).

6.4.5. Training materials

93. Pre-training planning includes preparing all training and fieldwork materials (e.g., introduction letters, household tracking lists, maps, cluster completion sheets, etc.), programming software, and testing hardware. In many cases, only the most essential documents need to be printed, and others can be made accessible online.

94. For computer-assisted surveys, each participant should have the opportunity to use the technology during training to ensure familiarity and proficiency. However, granting access to certain tools, such as a software platform, at the beginning of the training session may distract the trainees. Therefore, it is advisable to schedule a practical session using certain tools at a later stage in the training.

95. Typical materials usually include:

- Printed copies of the questionnaire: Allow trainees to follow question flow and take notes.
- Interviewer manual: Contains information on the survey's general purpose, instructions for conducting interviews, guidelines for recording outcomes or responses, detailed explanations of the questions, references to the methodology for recording answers, and details on hardware and software (if applicable). It covers the roles, responsibilities, and protocols for proper execution. In designing training, keep in mind that, while this manual serves as an essential reference during training, interviewers generally do not or cannot consult it mid-interview (Fowler and Mangione, 1990; Kutka et al., 2023; Iarossi, 2006).[2]

- Supervisor manual: Covers topics such as how to assign units to interviewers (from both methodological and technical perspectives), ensure the completeness and quality of questionnaires, monitor interviews/interviewers, request or record deviations from the survey design (e.g., a sample substitution), report and address unanticipated challenges in the field, and manage administrative matters relevant to the project.

- Video tutorials: Can be created for various modules covered in training sessions, for instruction on general practices (e.g., refusal conversion) and/or for specific training on a particular survey (e.g., how to code a specific question on a given questionnaire). This tool is particularly important when distance or blended learning methods are used (Daikeler and Bosnjak, 2023). The Malaysian Statistical Institute is one organization that uses video tutorials as a common training component. Professional video editing and production is ideal but may be daunting for low- and middle-income countries or organizations with limited budgets; however, free online software offers viable alternatives, and emerging artificial intelligence (AI) tools are expected to further reduce the time and cost of creating high-quality video content.

- Post-training assessments: Paper or electronic questionnaires that assess the effectiveness and satisfaction with training. Feedback is more likely to be more honest if it is collected anonymously. Areas of evaluation can vary, including the organization of the training, duration, instructional management, clarity of content, completeness of the topics covered, practical utility for the job, and overall satisfaction (Grassi and Romano, 2020). An example provided by Kutka et al. (2023) is a feedback form that asks participants to rate various dimensions such as the overall

[2] Daikeler and Bosnjak (2023) find that the use of supplementary training manuals improves response rates.

quality of the module, trainer performance, and their personal understanding of the materials. The form also offers several open-ended questions allowing participants to identify the most useful and most confusing parts of the training, as well as provide suggestions for future improvements.

- Evaluation test: Evaluates trainees' knowledge of study protocols and their ability to administer the questionnaire, ideally administered through both written quizzes and observation. In distance learning, tests can serve as checkpoints, as seen in the Italian Census, for which trainees had to pass intermediate tests to proceed but could retake them; such repetition can be an effective learning tool (Bali et al., 2023b). A final comprehensive, mandatory assessment should involve carefully designed questions and exercises. Additionally, interviewers should be observed during a "dress rehearsal," in which they simulate real-world interviews to demonstrate their readiness. In some cases, it may be appropriate to create a certificate to provide to interviewers who pass the training, and report on training attendance and on candidate training certification (including rates) (SRC, 2006).

96. Training materials should be translated into all languages used during data collection. When resources are limited, it may be feasible to translate only the most critical materials, such as the questionnaire and interviewer manual, or to prioritize the most spoken languages. If "on-the-fly" translations will be used in the field (e.g., for minority languages; see Chapter 4), this process needs to be covered in training to ensure a standardized implementation.

6.4.6. Training content

97. This section describes the content to include in training. The emphasis is on interviewer training, while also providing a brief overview of training specifically designed for trainers and supervisors.

6.4.6.1. Interviewer training

98. Interviewer training can be subsetted into three main types:

- *General interviewer training* (GIT): Designed to provide interviewers with foundational knowledge of their roles and standard rules for interviewing as well as practical guidance.
- *Refresher training:* Regularly occurring (e.g., every two years) training for previously-trained interviewers/teams; this is a slimmed-down and updated version of the general training.
- *Project-specific interviewer training*: Focuses on the needs of the specific project, typically examining the questionnaire, sample design, and protocols, as well as addressing challenges anticipated with the project (Stiegler and Biedinger, 2016; Sand et al., 2019).

99. *General interviewer training (GIT)* focuses on information and skills necessary for high quality interviews (Daikeler and Bosnjak, 2023). When working with external firms, it is important to assess whether the interviewers have sufficient level of general training; if not, GIT modules will need to be incorporated into the training plan.

100. GIT training sessions will typically cover the following:

- Recruitment and conversion: General practices for approaching potential participants and securing their consent to be surveyed
- Participant cooperation: Ways to tailor behaviour to the perceived features of the targeted individual and maintain interaction with them (Groves, Cialdini and Couper, 1992; Groves and McGonagle, 2001; see Section 6.6). General training on maintaining active cooperation might

provide guidance on how to keep a household engaged across time when the survey design involves multiple visits to the same household and/or might address how to engage with different populations (e.g., younger individuals, older adults, minority groups).

- Survey questions: General rules regarding the correct administration of the question-and-answer process, including transferring knowledge and skills regarding how to probe in a neutral and non-leading manner to reduce item non-response without introducing bias. While many projects require interviewers to repeat a question as worded when an interviewee seems confused or asks for clarification, others recommend a different approach. For example, Conrad and Schober (2020; see also 1999) argue that a Conversational Interviewing (CI) approach can be more effective; in this approach, interviewers clarify questions when they suspect respondents may have misunderstood. Even strong proponents of strict standardization, acknowledge that if the initial reading of a question does not yield an adequate response, then the interviewer ought to take some initiative to obtain the desired result (Fowler and Mangione 1990, p.35). Since these situations are dynamic and vary from one situation to another, the goal is to train interviewers who can effectively address these issues without introducing unintended semantic variations or influencing the interviewees (see also Fazzi et al., 2009; SRC 2011, par. X, p. 13). For more, see Chapter 4.

- Interviewer presentation: When relevant, training can provide guidance regarding the interviewer's attire. For example, a formal outfit worn by the interviewer can impact participants' willingness to engage (Guéguen and Jacob, 2015). In specific contexts, it is crucial to wear neutral clothing, as attire can convey particular attitudes or behaviours, leading respondents to give conformist or socially desirable answers. Even ordinary materials that interviews carry into the field can matter; in a study conducted in Nicaragua in 1990, researchers found that the use of pens that signaled differing political affiliations caused variation in survey responses (Bischoping and Schuman 1992).

- Ethics: It is important to train interviewers on ethics in survey research, including confidentiality issues and relevant regulations that apply to official surveys. Interviewers should be well-prepared to ask for informed consent at the beginning of the interview and to respond to requests for detail or clarification from the interviewee. In many countries, concerns about privacy have increased in recent years, so interviewers must be equipped with relevant training and support materials related to legal issues. Moreover, interviewers should receive training on gender norms and be aware of language that may not be inclusive. This awareness can vary significantly across different countries and contexts.

- Logistics: General guidelines related to logistics may include ways to minimize risks to interviewers in the field, how to handle emergencies, and how interviewers are monitored and supported in the field.

101. **Example:** In work by Paraguay's INE, the training sessions include information on the internal regulations of the INE, as well as clear expectations regarding each team member's responsibilities, acceptable clothing, required conduct, and appropriate behaviour, as well as the disciplinary sanctions for non-compliance with project protocols.

102. *Project-specific interviewer training* focuses on explaining details of a planned study, including project objectives, sample composition, sampling procedures, a detailed review of the questionnaire, use of the data, recording and transmitting the data, and other details. It is crucial that trainees thoroughly

understand and internalize underlying definitions and concepts referenced in the questionnaire. While Kutka et al. (2023) recommend focusing on the questionnaire section-by-section, trainers should spend more time on questions that are more ambiguous or challenging to answer. Especially for those difficult questions, trainees should practice categorizing ambiguous responses into predefined answer options, possibly using probing techniques.

103. Project-specific training should comprehensively address:

- Filter questions: Instruction on the conditions under which specific questions are asked, including the logic that routes respondents through the questionnaire
- Interdependencies: How certain questions relate to others within the questionnaire
- Project materials: Showcards, hardware and software (if applicable), etc.
- Survey mode: How to administer the mode(s) of the project (e.g., PAPI, CATI, CAPI)
- Contact attempts: Training on how to contact the household and/or potential respondent, create appointments (if applicable), register refusals, and determine household eligibility (if applicable)
- Guidance on electronic questionnaires (if relevant): Instruction that creates strong familiarity with the software and electronic questionnaire; how to enter responses; and hard and soft checks[3] implemented to ensure data quality, including how to interpret and address error messages to resolve issues flagged while data collection is in-progress.
- To the extent that some of the above is the same for all surveys conducted by an NSO, then various components can be moved to the GIT.

6.4.6.2. Training of trainers

104. In a TTT program, trainers are trained on all aspects necessary for successful data collection, to teach their interviewers what they learned at the centralized training. The TTT program should provide a model agenda for the small group training sessions. It is important to make sure there are enough trainers and that they have solid survey, context, and subject experience. Best practices for trainings discussed above apply to these trainings as well.

6.4.6.3. Training of supervisors

105. Given the important role that supervisors play in assisting with both project management and interviewer efficiency and effectiveness, it is important to conduct standalone training(s) for supervisors. This training should cover quality control, which will vary by project and mode (see section 6.8 of this Chapter on quality control). The supervisor should be also trained on standards for ethical and scientific conduct (SRC, 2011). The duration of supervisor trainings will depend on various factors but, for reference, Kutka et al. (2023) suggest a total duration of 2–3 days.

[3] Software for interviews can be programmed to include soft rules that function as alerts, but can be over-ridden to proceed with the interview – an example is a flag that notifies interviewers and/or respondents of inconsistencies in answers, allowing them the option to review and correct their responses. Hard rules, on the other hand, prevent the questionnaire from being continuing or being submitted until a given issue is resolved, ensuring data validity.

106. Supervisor training should take place outside of interviewer training hours. The supervisor should attend a substantial part of, or all, the interviewer training to ensure they are familiar with all aspects of the project and helps build synergy and team cohesiveness.

6.4.6.4. Other trainings

107. In addition to training activities that cover the roles of interviewer, trainer, and supervisor, projects need to plan and hold training sessions for those involved in production monitoring and quality control. Specific needs will vary by project, conditional on factors such as prior experience and knowledge; any use of new technology (e.g., new production monitoring dashboard or new software), and complexity of the processes.

6.5. Publicity of survey activity

108. Publicity for a survey can be targeted toward the public and/or authorities. Publicity targeted at the public is aimed at providing a general sense of project goals and activities to the public; it can be useful in building and maintaining trust in statistical organizations. Publicity targeted at authorities is aimed at securing support (e.g., endorsement) or approval (e.g., permission to enter a community) from authorities. Relevant authorities can be government employees, such as local officials, or unofficial leaders, such as gang members in areas controlled by criminal organizations.

109. Publicity is a highly context-dependent task. This section provides an overview of the different types of publicity that are sometimes required. Yet, the nature and scope of publicity efforts related to data collection will vary by the nature of the country and regional context, survey, local norms, and project budgets.

110. This section covers the three stages of publicity, including:

- Planning (Section 6.5.1)
- Testing (Section 6.5.2)
- Implementation (Section 6.5.3)

111. Within each subsection, the text describes publicity for both the public and regional authorities.

6.5.1. Planning

112. Planning for publicity needs to begin well in advance of data collection and involves three steps: identify goals, design messaging, and develop an implementation plan.

113. *Public publicity.* Publicity information needs to be relevant, understood, and accessible by the relevant group(s) targeted by the effort. Planning for public publicity involves three steps:

[1] Step 1: Identify goals by asking: Who will the publicity effort target and why? Relevant information may be gleaned from past experiences (e.g., are response rates low among a particular population subgroup?) and expert or community consultations (e.g., is the survey likely to encounter resistance?). The nature of the survey will also inform goals: for example, if the survey is targeted at the entire adult population, two goals might be to increase public confidence in the project and provide information about how to participate.

[2] Step 2: Design messaging by asking: What messages will be included in the publicity effort? In addition to past experience and expert consultations, one approach to generating and refining messaging is to conduct focus groups with community members. In these group sessions, different messages can be discussed to assess and select clear and compelling messages.

[3] Step 3: Develop an implementation plan by asking: How will the messages be circulated and what steps are required to prepare the materials? Publicity efforts may include brochures, posters, visual presentations (live and/or recorded and available online), distribution of a publicity centre phone number, press releases, social media, and websites that might be accessed via URL or QR code. For example, a media-centred approach focused on increasing awareness might involve collaborating with the press and using social media to generate news stories that, inform the public about the survey and its importance. Another approach might be materials-centred: for example, print posters and banners that can be placed in particular locales if the goal is to increase awareness of the survey in a particular region or community. In multi-language contexts, teams should allocate effort and resources to translation. Scrupulous proofreading of publicity material is important to ensure that efforts to increase support do not backfire.

114. *Authority publicity.* In some cases, publicizing the survey may be formally mandated by legal authorities and/or informally required by community leaders or other informal authorities. Planning follows the same three steps as for public publicity:

[1] Step 1: Identify goals by asking: Who will the publicity effort target and why? For example, in some countries, it is important to provide advance notice of a survey activity to local authorities. This may be legally required (e.g., [[EXAMPLE-PENDING]]) or a useful courtesy communication (e.g., [[EXAMPLE-PENDING]]). In the case of face-to-face interviewing across households in insecure contexts, members of the public or others may report data collection teams for what appears to be suspicious behaviour; if authorities have been notified of the team's presence, they can more easily handle such reports. In other cases, it may be useful to inform non-government authorities (e.g., religious leaders, gang leaders) of a survey to gain their support and/or permission to enter a particular community.

[2] Step 2: Design messaging by asking: What messages will be included in the publicity effort? Past experiences, expert consultations, and outreach to the relevant leaders can be effective ways to determine what approach to messaging is required and/or will be most effective.

[3] Step 3: Develop an implementation plan by asking: How will the messages be circulated and what steps are required to prepare the materials? The channels used for messaging will vary significantly: e.g., when advance notice is targeted at government officials, phone calls or text-based messaging may be appropriate. While some publicity efforts can be conducted in advance of fieldwork, other cases may call for members of a face-to-face fieldwork team to carry letters of introduction with them, to be handed over to local authorities on entering a given area. As with publicizing activities aimed at the general public, it is important that messages are in the appropriate language(s), contain relevant information, and are professionally formatted (e.g., on letterhead, signed).

6.5.2. Testing

115. It is useful to test publicity activities to ensure they achieve their goals. There are two types of testing:

[1] **Pretesting**. The pretesting of publicity materials begins during the planning phase, when particular messages (and/or their design) are discussed with experts, community members, or relevant leaders. Whether through focus groups and/or a series of in-depth interviews, the goal is to collect reactions to the materials and then refine them. In some cases, community leaders who are brought into the pretesting process may end up supporting the data collection effort, for example by providing perspective on how to reach and recruit the intended sample.

[2] **Piloting.** Once publicity messages and channels are determined, it is can be useful to launch a pilot test – reaching a subset of the intended targets – in order to observe whether there are any unexpected challenges (e.g., unexpected reactions or lack of response to messages or problems with QR codes or websites (if applicable).

6.5.3. Implementation

116. Publicity survey activities will be implemented in several ways, depending on the particular goals and target population.

117. Publicity for the general public might be implemented via messages posted on billboards, posters, leaflets, newspaper ads, tv spots, radio spots, or the like. Websites can serve a legitimating role, by providing more information about the organization, past survey projects, and the current project (e.g., a timeline of activities and plans for post-data collection reports or other output).

118. Some projects distribute information via WhatsApp or similar platforms. This may involve conveying information to contacts (if available) in each community so that information is dispersed in stages via social networks. Or it may be possible to distribute QR codes or SMS messages to publicize a survey activity.

119. If the goal is to socialize local authorities into understanding and supporting the project, then the approach might involve individual or small group meetings with relevant people who have influence in the places where data collection is scheduled. If the goal is to gain written permission from local authorities, then implementation should focus on issuing these requests according to context-specific rules and norms.

120. **Example**. In the implementation of Indonesia's Nutrition Status Survey, which was carried out by the Indonesian Ministry of Health, notification of activities was carried out in stages. A first step was to generate material that described the aims and objectives of the survey, the target population, and the hope for support from local residents and authorities. Then, a first stage began with outreach to officials in each province through meetings and discussions. The next stage involved outreach from the provincial level to the district/city level. A component of this activity involved obtaining verbal permission from village heads, which facilitates the data collection team's access and safety in these locations. As a next stage, the village head notifies neighbourhood heads, using project brochures. The plan also called for socialization at the regional level, which was carried out by the local health department.

6.6. Data collection implementation

121. In addressing data collection implementation, this section focuses on in-person interviewing (PAPI, CAPI), but some sections are relevant to CATI and other interviewer-administered approaches. Other modes of data collection (e.g., self-administered surveys conducted via mail or the web) mostly have different requirements and considerations when it comes to data collection implementation (see, e.g., the ESS, 2024 on modes of data collection); this section does not address data collection for those modes.

122. This section covers three core stages of the data collection process:

- Approach strategy – how to make initial contact (Section 6.6.1)
- Recruiting participants – how to request participation (Section 6.6.2)
- Interviewing – how to conduct the survey (Section 6.6.3)

123. Inefficiencies and error can be introduced at any point in these stages and, thus, planning for data collection requires attention to all three stages. The goal of this section is to advise on how to develop

comprehensive strategies for approach and recruitment, and to present guidelines for best practices for interviewing.

6.6.1. Approach strategy

124. The first stage in data collection is approaching potential respondents. This section discusses three commonly considered elements:

- Incentives
- Advance and reminder letters
- Contact logistics

125. The relevance, combination, and composition of these elements and how they constitute the overall approach strategy varies across and within projects. It is important to recognize that countries will differ in their ability to utilize certain approaches or elements of them. For example, whether advance and reminder letters are appropriate depends on both the context (literacy of the survey population, existence of a functional postal service, etc.) and the survey mode. Thus, practical constraints (context, mode) are important to keep in mind alongside the goals of efficiency and effectiveness (error reduction), while adhering to best practices in the ethical treatment of human subjects.

6.6.1.1. Incentives

126. Incentives are tokens of appreciation and/or compensation for the time it will take someone to participate in a survey (Pforr, 2016; Stoop et al., 2010). Prior to incorporating any incentives, it is important to consider these questions (Singer et al. 1999):

- What is appropriate within the context (e.g., local norms)?
- What is reasonable within the budget? In answering this second question, keep in mind that the use of incentives may end up being less costly than using no incentives if, for example, the choice of no incentives results having to send more reminder letters, make more interviewer visits, and/or extend the data collection period (Lipps et al, 2019; Lynn et al. 1998).
- Will incentives exacerbate representation bias (e.g., increasing participation by types who already participate in greater proportions)?
- What effect will incentives have on long-term expectations among the targeted population (e.g., some local firms may end up priced out of the survey market if the public expects incentives for survey participation)?

127. The discussion in this section is for situations when incentives are possible and appropriate. Incentives can be conditional on participation or non-conditional (provided to all sample members, even if they do not participate) and monetary or non-monetary. To develop the optimal incentive strategy given the trade-off between survey quality and time and budget constraints, one needs to consider what approach – or combination of approaches – will work best for a particular survey in a particular context. In addition to relying on past experiences and expert consultations, it can be prudent to run tests before committing to a given incentive strategy.

Conditional incentives or unconditional incentives

128. Some research finds that unconditional incentives are more effective than conditional incentives when it comes to increasing response rates, and both tend to be more effective than no incentive at all (see, e.g., Groves and Couper, 1998; Stoop, 2005; Singer and Ye, 2013; Kumar et al, 2022). Higher response

rates indicate more success in realizing the intended sample, yet high response ratees only decrease nonresponse bias if they are achieved by recruiting in types who would otherwise not participate.

129. Another approach is to use conditional incentives – that is, incentives that are only provided if the individual participates. In such cases, it may be possible to use an advance letter (see next subsection) to make the incentive salient, which may make conditional incentives comparatively effective while saving on costs (Lasky-Fink and Rogers, 2020). In other cases, it may be efficient (resource-saving) and effective to use a combination of both conditional and unconditional incentives.

Monetary or non-monetary incentives

130. A good deal of research, albeit much of it focused on high-income countries, finds that monetary incentives are more effective in increasing response rates compared to non-monetary incentives, and non-monetary incentives are more effective than no incentive (Lipps et al, 2019; see also Edwards et al, 2005). Less research focuses on reduction in nonresponse bias. For modest amounts, the higher the monetary incentive, the higher the response. However, there is a limit: very large amounts may cause response rates to decrease (citation pending). Furthermore, extremely high monetary incentives can be perceived as untrustworthy or, in the case of official surveys, as a waste of public money; they can also be unethical, if they compel a person to participate out of need. The decision over what amount to offer should be informed by knowledge about *and* consistent with local customs and only considered when it is not seen as illegal, offensive, or coercive.

131. Some examples of non-monetary incentives are the following: a key chain; a pen; a lottery entry to win a costly item or actual money; and a voucher that is commonly accepted in many local stores to use to buy food items. It may be that non-monetary incentives that are more akin to actual money are more effective. As noted, it is important to consider the target population and how they will appreciate this non-monetary incentive (local customs) and if there are groups that may be more or less inclined to participate as a result of this, given that differential responsiveness to incentives can increase non-response error.

Practical and legal restrictions

132. Some issues to consider include constraints (if any) on what can be delivered legally (e.g., is it allowed to provide monetary incentives), safely (e.g., can an interviewer carry monetary incentives with them without risks), and reliably (e.g., will the incentive in the envelope be sufficiently noticeable to the addressee to avoid being thrown away). In addition, it is important to have a plan for how to monitor incentives in the field; at a minimum, that monitoring should ensure that incentives delivered by data collection team members reach the intended individuals and, where possible, such monitoring should collect data that permits an assessment of how effective the incentives are, and for whom.

6.6.1.2. Advance and reminder letters

133. Advance letters (also known as prenotification letters or invitation letters) inform the potential respondent about the survey prior to contacting them for an interview (Dillman et al, 2014; Stoop et al., 2010). Advance letters can be used for the general population or for specific target subgroups, particularly if these subgroups are known to have lower response rates. Advance letters can increase response rates, although they may be less effective when the letter is addressed to the household (Capistrano and Creighton, 2022). Reminder letters are sent to potential respondents who have been contacted but have not yet completed the survey.

134. The content in a typical advance (and reminder) letter addresses two goals: recruitment and informed consent. With respect to recruitment, the letter should address motivations and/or barriers to participation. The former could focus on the benefits of participating, while the latter addresses potential perceived costs

or concerns (see e.g., Stoop et al., 2010). With respect to informed consent, ethical and legal guidelines in human research require providing sufficient information for an individual to make a reasonable decision about whether or not to participate. Advance letters can be used to provide some of this information. For example, General Data Protection Regulation (GDPR) rules, surveys in European Union countries must include a data protection statement or a link to it in the advance letter.

135. **Box 6.1** indicates elements that typically appear in an advance letter.

Box 6.1. Content to include in an advance letter

PURPOSE*:* A statement that conveys the goal of data collection (e.g., a census, research)

TOPIC*:* Information on (some of) the survey topic(s)

EFFORT*:* Amount of effort (e.g., time) the survey requires

CONFIDENTIALITY AND DATA PRIVACY: An explanation of how confidentiality will be safeguarded, a data protection statement, information about participation rights (e.g., right to withdraw); individuals' (data subject) rights (mandatory in some contexts), and reference to relevant rules and regulations (if applicable)

SELECTION: A short statement explaining why this person has been selected and that participation is voluntary

CONTACT: The contact details of the NSO (or the subsidiary if another survey organization conducts the data collection), either in text and/or via a QR code

INSTITUTION: Letterhead, logos, or QR codes that link to the relevant institution to convey legitimacy as well as the signature of a respected, named individual, such as the NSI director or study PI

MOTIVATION: Rationale for participation (e.g., to provide their voice in a process); benefits from participation and/or information on conditional or unconditional incentives (if applicable)

RISKS: A statement on any potential risks to participation

PROCESS: Information on next steps regarding participation; if applicable, information on how to initiate participation (e.g., a survey link or contact information for scheduling an interview)

136. In designing an optimal advance and/or reminder letter strategy, some important considerations are implied vs. explicit consent, languages, addresses, visual design, and logistics (see, e.g., Dillman et al., 2014 or De Jong, 2016). These topics are discussed below.

Implied or explicit consent

137. General principles regarding the ethical treatment of human subjects, as well as many countries' rules and regulations, require informed consent from a potential respondent to participate. *Implied consent* is when an individual is informed about the nature of a survey and chooses to participate; although they have not provided a signature or otherwise recorded their consent in an *explicit* fashion, the act of participation suffices to demonstrate that they agreed to participate. Implied consent may be reasonable when data privacy is high and survey content is not particularly sensitive. In these cases, if an advance letter provides sufficient information for an informed choice (see above Figure 6.1), some view a subsequent decision to participate as implied consent. In these cases, it is important to ensure the participant received the advance letter and the letter should state that participation is understood to indicate consent; when there is uncertainty, relevant information should be revisited when direct contact is made.

Languages

138. In multi-language contexts, it is prudent to include different language versions of the advance and reminder letters or to summarize the important elements in the most common languages. It is also important to indicate how to participate in the preferred language.

Addressee

139. In general, advance and reminder letters have a greater chance of being opened and being effective if the name of the sampled person is known (in case of a named person sample) (see, e.g., Capistrano and Creighton, 2022). Who is likely to open the letter also has implications for the incentive strategy. In case of a household or addressed-based sample, it is important to use the correct salutation based on eligibility criteria and other important country-specific customs.

Visual Design

140. The choice of envelope as well as the layout of the letter matters. Dillman et al (2014) have written extensively about the effect different types, sizes and colours can have on the likelihood of opening the envelope and reading its content by potential respondents.

Logistics

141. It is important to carefully plan logistics around advance and reminder letters to maximize efficiency and effectiveness. This means deciding upon how many letters to send to a sample unit; the duration between sending the letter and contacting the sample unit; how much time to leave between subsequent letters; what days of the week to send/deliver the letters (Lynn et al. 2024); how the letters will be delivered (e.g., postal, by interviewer, etc.), and to whom (e.g., to individuals, households or communities); how to track the status of advance and reminder letters (e.g., at the sampled unit level); and how to actually carry out (e.g., print, stamp) and budget (e.g., mailing costs) for these activities.

Additional considerations for reminder letters

142. First, consider providing new information (e.g., highlight different survey topics) to motivate people to participate, given that these nonrespondents were not swayed by the content of the advance letter. Second, reminder letters can include other options that may enhance participation, such as the inclusion of a self-completion postal questionnaire as part of a sequential mixed mode design (see Chapter 5); the inclusion of different unconditional incentives; or information about a higher valued conditional incentive if a mixed incentive strategy is part of the overall approach strategy. Third, it can be useful to consider a different envelope, different layout of the letter, or different type of communication such as a postcard (see Dillman et al, 2014). Information learned from prior studies and/or from response rates to advance letters can inform decisions over a reminder strategy; for example, if data collected from responses to advance letters shows varying propensities to participate across subgroups, one could tailor the reminder strategy to target particular subgroups with novel messages.

6.6.1.3. Contact logistics

143. The contact strategy concerns the how, when, and how often a sampled unit (person, household or address) should be approached by the interviewer. In some cases, the contact strategy involves determining how an interviewer will select a respondent: for example, when using a random route/walk approach, the interviewer must follow protocols established in the sample design (Chapter 5) and covered in the project-specific training (see Section 6.4.1) to determine which household to approach. Once interviewers have made contact, their effort turns to recruiting individuals to participate (see next section).

144. Below are several considerations relevant to developing a contact strategy:

- *Mode of contact attempt.* Most projects will use the same mode for contact as for the actual interview. For example, in-person interviews will make contact attempts at the location of a sampled unit. Certain factors can affect contact success for in-person approaches, including locked buildings and gated communities. Phone-based surveys will make contact by dialing phone numbers of potential respondents; software can be used to make (e.g., using random digit dial) and transfer these calls to interviewers. It is also possible to mix modes across contact attempt and interview: for example, as indicated above, advance letters can be mailed to potential participants, who then act on the information in that letter to set up a phone or in-person interview. Reaching out by phone or text to set up an in-person interview is possible when there is a high match between phone numbers and unit locations (addresses); yet, even in rare cases when that match exists, reaching out via phone or text could lead to higher levels of refusal because many people find it easy to refuse and/or are put off by this type of approach strategy (Kappelhof, 2015b, 2017).

- *Interviewer assigned to contact attempt.* In some cases, there is a need to match gender, language, ethnicity of interviewers to potential respondents (Kappelhof, 2017); this needs to be considered when developing the approach strategy.

- *Interviewer involved in selection.* Especially in the case of in-person surveys, interviewers may play a role in sample design by selecting households. This is the case for random route/walk approaches and/or an interviewer might be required to first list the eligible households in case of multiple households at a single location and subsequently select and contact the correct household via a predetermined procedure (see Chapter 5). Interviewers may also need to select the correct person to be interviewed from the eligible household members via a predetermined selection procedure (e.g., identifying the head of household or selection via a household screener such as a simple random sample, first- or last-birthday, Kish grid, or other method; see Chapter 5).

- To the extent to interviewer is involved in selecting the household and/or potential respondent, the contact strategy needs to include clear instructions and training as well as monitoring to ensure the correct person is approached. For example, to assess eligibility, interviewers need to know who should be included as a member of the household and as eligible – e.g., whether domestic workers count as members of the household and, for eligibility, whether to count household members who are away from home for longer periods of time (e.g., students or seasonal workers). Interviewers need to know what constitutes a household in case of a survey aimed at people living in private households – for example, how to determine if a student house is a household or how to consider people living in temporary housing. Finally, interviewers need to know whether a proxy interview is allowed (i.e., another family member can be approached to participate in the interview on behalf of sampled person).

- *Time of contact.* Interviewers' working hours should align with the (likely) availability of sampled persons (Beullens et al., 2018). Different subgroups of the population have different at-home patterns (e.g., working people vs. retirees). The day of the week and the moment of the day can affect successful contact rates overall and for different subgroups (see, e.g., Stoop, 2005; Stoop et al., 2010). To minimize nonresponse bias, a contact strategy should vary the day and hour of contact attempts, especially when recontacting those who were unavailable on the first attempt. For example, for in-person surveys, the European Social Survey recommends making at least contact attempt during the day, one in the evening, and one on the weekend (Stoop et al., 2010; ESS, 2024).

- *Number of contact attempts.* More contact attempts can reduce non-response errors and improve representation but will require additional time and budget. As such, the exact number and optimal allocation of the number of contact attempts is dependent on the budget, team capacity, and schedule. Some strategies to boost efficiency are to prioritize recontact attempts for units that were not unavailable at the time of the first contact attempt and, for in-person interviews, to equip the interviewers with information (e.g., advance letters, calling cards) to leave at the home of the non-contacted sampled unit.

- *Recording contact attempts*. Contact attempt outcomes must be recorded as paradata. While the contact strategy is being implemented, temporary disposition codes can be entered (e.g., contact made and interview scheduled). Ultimately all contact attempts need to have a final disposition code (see AAPOR 2023). Collecting and assessing data on contact attempts facilitates responsive (Groves and Heeringa 2006) and adaptive survey design (Schouten et al., 2018). For example, one could distribute contact attempts differentially across the sample based on varying (estimated) response propensities for population subgroups (see, e.g., Calinescu et al., 2013; Schouten et al., 2017).

- *Recontacting.* It is important to pre-plan how to handle noncontact; in some multi-stage designs, for example, unsuccessful contact cases from a first phase will be assigned to another interviewer in a second phase. Recontacting also comes into consideration in cases in which a contact is made but the target person or household declines; the next section addresses these situations in discussing recruitment strategies.

145. The aim is to develop a contact strategy that is resource-efficient and that minimizes nonresponse error. For in-person surveys, common challenges to anticipate are buildings with multiple households with a centrally controlled access point or gated communities where, in order to make contact, the interviewer first needs to gain access. Interviewers need to be aware of these situations and trained to deal with them effectively. More generally, research has shown that, on average, certain socio-demographic groups have lower rates of contactability, such as people living in large urban areas, people who are single, people with jobs that require them to be away from home more often, apartment dwellers, people in a higher socio-economic status group and (depending on survey design choices) minority groups albeit for different reasons (Groves and Couper, 1998; Stoop, 2005; Kappelhof, 2015a; Beullens et al., 2018). Conversely, research has shown that the elderly, larger households, especially ones with younger children are more likely to be at home and therefore easier to contact (Groves and Couper, 1998; Stoop, 2005; Beullens et al., 2018). Information about contactability patterns among the targeted sample for a given project can be taken into consideration when developing a contact strategy that minimizes the extent to which certain groups in the target sample are over- or under-represented.

6.6.2. Recruiting participants

146. After contact is made with the household or sampled person, the recruitment stage begins. There are interdependencies between the approach strategy stage (Section 6.6.1) and the recruitment stage. As discussed below, when a recruitment attempt fails, decisions need to be made about whether to re-contact the person, how, and when. While those topics are covered in the prior section, some of them are revisited below in discussing unsuccessful recruitment outcomes.

147. The aim is recruiting participants is to realize the intended sample in a way that adheres to project protocols and ethical standards, while also minimizing nonresponse error. To engage in participant recruitment, an interviewer should be aware of project protocols and ethical principles. This includes knowing, among other things:

- how to introduce the project and secure informed consent
- when and how to issue an incentive (if applicable)
- whether or not substitution is allowed
- whether or not it is permissible to include family members or other persons present (for example when the sampled person does not understand the language and a family member translates)

148. To minimize the risk of nonresponse bias, an interviewer should also be aware of factors the contribute to the risk of non-cooperation and know how to minimize the likelihood of an initial refusal. This includes, among other things:

- Familiarity with local customs for addressing individuals or, for in-person surveys, for entering a house (e.g., whether interviewers need permission from local leaders or other members of the household [see, e.g., De Jong, 2016]).
- How to establish a positive rapport. In particular, interviewers should be trained to be empathetic and understanding of different types of people and their circumstances. Their choice of words and nature of interaction should make the target respondent feel comfortable.

149. As noted previously, in some cases, rapport alone will not be sufficient – for example, in some cases, it may be necessary to match the gender of the interviewer and the potential respondent or household informant in case of the screener phase (see also the earlier section on refusal conversion).

150. When contact is made but the target person declines, that unit receives the fieldwork disposition code refusal (or in case of a household refusal a refusal by proxy). Whether this is the final disposition or a temporary disposition depends on whether the protocols permit re-contacting refusal cases. Research has shown refusal conversion can reduce the potential for nonresponse error and improve upon the net sample composition, but success varies on effort and skill so it is advisable to invest in trainer interviewers on refusal conversion efforts (see, e.g., Beullens et al., 2018; Stoop et al., 2010). The development of an effective and efficient refusal conversion strategy can be helped by considering several important elements such as how to re-contact refusals, who among them to re-contact, how many of them to re-contact, and when to re-contact them. Stoop et al. (2018) also provide a list of some of the most common reasons for refusal; by integrating this into the data collection plan, these reasons can be systematically recorded by the interviewer.

151. The reasons people give when refusing to participate can be used to develop different and more effective arguments or make other relevant changes to the approach strategy. For example, if reasons for refusals highlight needing a gender or ethnic match, one could consider sending another interviewer that meets those requirements. Or, in the case of refusal reasons such as ‘no time’ or ‘too much effort,’ one could assume this is situational and callback at another time, or increase the incentive (if budgeted) and emphasize this as an acknowledgment of how much their time is valued, or point to other benefits of participating (Dillman et al. 2014; Groves and Couper, 1998; Stoop, 2005; Stoop et al., 2010).

152. When it comes to who to re-contact, it is typically more efficient to focus on sampled units that are classified as *soft* refusals versus *hard* refusals. *Soft* refusals are the sampled persons who did not flat out say that they do not want to participate, but rather use some sort of excuse, such as no time or no interest in this topic. Recontacting hard refusals is be less efficient, but the data collected on hard refusals maybe useful to identify whether the hard refusals seem to mostly come from a particular subgroup in the sample (e.g., people with low trust in institutions) or because of a particular reason (e.g., lack of gender or ethnic match to the interviewer). In instances when there is an expected correlation between the reasons not to

participate and someone's view, attitude, background, type, etc. (i.e., omission of the unit increases likely nonresponse bias), targeted efforts around refusal conversion could be help reduce nonresponse error.

6.6.3. Interviews

153. Once an individual has agreed to participate, the interviewer must administer the interview according to protocols. Low quality interviews can increase overall nonresponse (e.g., via early termination such that the interview has to be cancelled), item nonresponse (refusal to answer a question[s]), and measurement error. This section provides tips on how to conduct high quality interviews.

154. **Box 6.2** provides guidance for high quality interviews. These practices originate with the interviewer and should be emphasized in training (see Section 6.4). In addition, it is useful to note that respondent behaviour can contribute to low quality interviews. Some common behaviours by respondents, especially reluctant ones, are acquiescing and satisficing, although there are additional forms of response style behaviours that introduce variance and/or error (see. e.g., Tourangeau and Rasinski, 1988; Krosnick et. Al, 1996; Billiet and McClendon, 2000; Morren et al., 2012). These behaviours can be the result of a respondent having low capacity (an issue magnified by task difficulty, for cognitively burdensome questions) or, instead, a motivation to provide an inaccurate response, such as the tendency to provide socially desirable answers and/or to censor (not disclose) on topics that are sensitive (Hui and Triandis, 1989; Kappelhof and De Leeuw, 2017). It is important that interviewers are able to detect issues and, to the extent possible, keep respondents motivated and focused throughout the interview; such efforts reduce the potential for measurement error as well as minimize break offs.

155. When sensitive topics are included, interviewers should be prepared to take appropriate actions, such as terminating the interview early and/or referring the respondent to resources that can assist the respondent through personal challenges. When sensitive topics relate to criminal acts or abusive issues, not only is respondent censoring a potential issue but, as well, there could be legal consequences to the disclosure of these behaviours. Interviewers must know how to present such issues, must be aware of legal implications to the disclosure of such behaviours, and must be ready to respond appropriately if a respondent requests help for a certain issue or situation.

Box 6.2. High quality interviewers do the following

- Adhere to the question ordering
- Read questions (and response options) verbatim reading of questions
- Present show cards efficiently and correctly (if applicable)
- Speak clearly and at a good volume and pace
- Do not interpret questions (unless permitted by protocols)
- Use neutral probes, when necessary (e.g., ask respondent to provide an answer that matches a response category)
- Do not translate to another language on-the-spot (unless permitted by protocols)
- Record all responses accurately
- Record open-ended answers verbatim
- Use the same, standardized approach across interviews
- Maintain consistency in offering a "don't know" option (following protocols)
- Avoid value-laden or visible reactions to responses (so, "uh-huh" instead of "yes, good")
- Avoid and/or record the presence and influence of bystanders (for an analysis of bystanders' influence within the Afrobarometer, see Zimbalist 2022).
- Adopt a neutral physical appearance

156. Measurement error can also be affected by other elements during the interview, such as interviewer attributes (e.g., race or ethnicity of the interviewer), interviewer-interviewee interaction (e.g., gender match), or the presence of influential third parties / bystanders (Holbrook et al., 2019; Kappelhof and De Leeuw, 2019). These are not necessarily under the control of the interviewer during the interview process, but may be important factors for interviewers to be aware of and/or may be factors on which data should be recorded for later analysis (e.g., analyses that assess interviewer effects by looking at correlations between responses and interviewer gender or ethnicity).

157. At the end of an interview, the interviewer should state the interview is over, thank the respondent for their time and effort, and provide the incentive (if applicable). They should answer additional questions from the respondent about the survey, such as when and where study information will be available. The interviewer should close the survey (or for PAPI, write the end time and any other closing information on the paper form. Furthermore, (if relevant) the interviewer may complete a short interviewer assessment about the interviewee's characteristics, their own characteristics (e.g., gender, ethnicity [if not already collected in a format that can be linked back to that particular interview]) and/or the overall quality of the interview. For example, some projects (e.g., ESS, AmericasBarometer) require that, upon leaving the respondent, the interviewer fills in a short questionnaire that includes questions such as how difficult the respondent found the questions and whether there were bystanders present during the interview.

6.7. Production monitoring

158. Production monitoring involves the collection and assessment of actionable information regarding the unfolding of data collection, which permits statistically informed evaluation and managing of the survey (Kreuter et al., 2010). In simple terms, production monitoring involves continuous observation of survey processes to ensure that they are proceeding as planned. Because production monitoring is designed to permit evaluations of and changes to the data collection effort while it is in progress, it is particularly useful in the case of adaptive and responsive designs (see later discussion in this section).

159. Production monitoring is essential for detecting and addressing issues or deviations from the original survey design, and ultimately meeting the goals of data collection efficiency and efficacy. To successfully meet these macro-objectives, specific sub-objectives must be achieved. For example, for face-to-face and telephone interviews, it is essential to promptly verify that each sample workload is allocated in a timely manner to each interviewer, as well as to evaluate interviewers' progress and the quality of the responses they are obtaining (Wuyts and Loosveldt, 2020; Goel et al., 2021).

160. Advance planning for production monitoring – and responses to these efforts – is critical. To effectively address issues when they arise, improvement and communication strategies must be pre-established and available for implementation. For example, strategies may include coaching interviewers who are underperforming and providing additional training or support as needed. If necessary, further actions such as reallocating workloads to different interviewers or emergency recruitment in a specific region may be required (Walsh and Coombs, 2014).

161. Production monitoring (this section) and quality control (Section 6.8) are both essential for ensuring data quality (United Nations, 1984; Giuliani et al., 2004; De Leeuw et al., 2008), but they serve different purposes and may be applied at different stages of the survey process. Production monitoring entails the ongoing observation of the survey process to verify it is aligned with established plans and objectives, and necessitates the implementation of adjustments in real-time. Quality control involves systematic procedures to ensure that the data collected meets the pre-established standards for accuracy and validity; quality control can take place both during and after fieldwork. As such, there is natural overlap between production monitoring and quality control efforts. To make a distinction that will be useful in distinguishing the two,

this chapter considers production monitoring to centre on tracking progress against benchmarks, recording deviations from the original design as a form of paradata, and identifying patterns or trends that might signal emerging issues.

162. This section addresses the following topics:

- Planning (Section 6.7.1)
- Performance indicators (6.7.2)
- Design (Section 6.7.3)
- Monitoring (Section 6.7.4)
- Dynamic, responsive, and adaptive design (Section 6.7.5)
- Report (6.7.6)

6.7.1. Planning

163. Prior to initiating data collection, it is important to establish objectives and procedures for monitoring data collection. This should involve the following steps (Survey Research Center, 2010):

[1] *Develop production monitoring goals*: Specify monitoring objectives and which aspects of the process will be observed (e.g., progress and pace in data collection overall and/or by interviewer)

[2] *Identify a set of performance indicators*: Choose specific metrics that will permit assessments of progress toward selected objectives. Indicators could include both quantitative measures (e.g., completion rates) and qualitative measures (e.g., stakeholder satisfaction)

[3] *Develop monitoring tools and techniques*: Define tools and systems to monitor and visualize process performance (e.g., dashboards and graphs)

[4] *Developing an improvement strategy*: Formulate a strategy for incorporating feedback from monitoring activities into adjustments via communication and interventions

[5] *Develop templates for documentation and reporting*: Establish a system for regular reporting to project leaders and (if applicable) stakeholders on the status of the survey, including any issues identified and corrective actions taken

[6] *Create processes for evaluation and revision*: Regularly assess the effectiveness of the monitoring plan and revise it as necessary to adapt to changing circumstances or new insights.

6.7.2. Performance indicators

164. This discussion will focus primarily on production monitoring indicators that are quantitative in nature; these will be referred to as Key Performance Indicators (KPIs).[4]

165. With technological advances, it is possible to collect large volumes of paradata (also known as auxiliary data) while data collection is in progress, in a nearly passive manner (West, 2011). Electronically collected paradata is possible across various survey modes, including CAPI, CATI, and web surveys, and even mail surveys, when they are associated with electronic (postal) service logs. Electronic systems

[4] It is important to recognize that some indicators that are not strictly "key" – in the sense of being critical to the success of the survey – may also be included in a production monitoring plan and, further, some projects may benefit from including qualitative components (e.g., direct observation of interviews, open-ended questions to team members).

generate paradata as by-products of the survey process, such as call records, response rates, interviewer observations, time stamps, interview duration, geolocation, keystroke data, travel and expense records, and other related information (Kreuter et al., 2010). Copious amounts of paradata are possible but can be counterproductive when they overwhelm analysis and decision-making. Therefore, it is crucial that the collection and assessment of paradata are driven by specific goals. That is, paradata should be collected with a well-defined purpose (Kreuter et al., 2010; West, 2011).

166. A vast array of KPIs can be derived from accessible paradata. The selection of KPIs should be limited to those that are the most effective in highlighting anomalies in fieldwork (Fazzi et al., 2022; Jans et al. 2013). Moreover, KPIs should be objective and consider international standards (Giuliani et al., 2004) (see UN Official Statistics Principle 2 in Chapter 1). The European Statistical System has developed a handbook on enhancing quality through paradata analysis, which may provide some projects with a starting point. However, each project needs to identify the most relevant process variables tailored to its unique context (Kreuter et al., 2010).

167. **Table 6.1** lists several potential KPIs (West, 2011; Fazzi et al., 2022; Giuliano et al., 2004).

Table 6.1. Examples of KPIs for survey production monitoring.

KPI	Definition
Response rate	The number of interviews divided by the number of eligible (and/or eligible + unknown eligibility) sampling units
Activity rate	The number of units with a final outcome divided by the number of contacted units
Cooperation rate	Number of interviews divided by all the contacted eligible units
Refusal rate	Refusals (and could include breakoff cases) divided by all the eligible units
Eligibility rate	The number of eligible units divided by the number of total (eligible plus ineligible) units in the sample
Replacement rate	The ratio of the number of replaced households or individuals to the total number of households or individuals originally selected for the survey
Non-interview rate	The number of units for which the interview could not be carried out divided by the number of units with a final outcome
Item non-response rate	Number of “don’t know”/no answers divided by the total number of responses given to a question
Sample coverage	The proportion of units allocated to the field team that have received a final disposition (outcome) code
Interview time	Length of interview, from start to stop; can be assessed overall (e.g., to assess whether projections regarding questionnaire time were accurate) or per interviewer

168. KPIs should be reported at regular time intervals during the fieldwork period. It is important to note that measures can be assessed at different levels, such as overall survey, region or area, and individual interviewer. When calculating KPIs for individual interviewers, it may be important to account for the

nature of their workload; for example, if household size affects interview time and is unevenly distributed across interviewers, then calculations of average interview time should be adjusted for household size,

169. Intervals for checking and levels at which checks are performed need to be established in advance. Some measures may be relevant at all levels, while others may only apply at certain levels. In the 2022 census of Non-Profit Institutions (NPIs) conducted in Italy by the National Statistical Institute (Istat), KPIs are produced for each interviewer and by province, to determine whether problems are common to all the operators working in specific areas of the country or if they are limited to certain interviewers only (Fazzi et al., 2022). In terms of intervals, KPIs can be produced based on the day of the week or the hour, allowing evaluation of which days or times are most productive; this type of information can be used to tailor current or future interviewer assignments in ways that gain efficiency.

170. Given that both production monitoring and quality control rely on paradata, an important part of the planning process is identifying which paradata are needed for which purposes. It is often impractical to have a perfectly clear delineation of the two efforts – production monitoring and quality control. The below example illustrates production monitoring work that also has implications for goals related to quality control.

171. ***Example***: In ISTAT's CAPI/CATI Labour Force Survey, a key tool for monitoring field performance is the transmission report. Interviewers are instructed to connect to the web at least once daily to retrieve the names of the sampled households and submit interview outcomes. By analysing the number of connection attempts, transmission attempts, and successful connections, it is possible to identify whether a data collection issue for a particular interviewer stems from transmission difficulties or from delays in conducting the interviews (Giuliano et al., 2004).

6.7.3. Design

172. This section addresses how to design systems for measuring and displaying production indicators. The first step is to create a system that can record and process paradata to obtain KPIs. The second step is to visualize these indicators. Both steps need to be designed in advance of the start of data collection.

6.7.3.1. Step #1: Recording and processing paradata

173. Paradata can be automatically generated (e.g., keystrokes) or actively collected (e.g., by interviewers). Increasingly sophisticated software permits full automation in ways that reduce burden on interviewers and decrease risk of accidental or intentional misreporting of paradata (Schenk and Reuss, 2024).

174. For actively collected paradata, it is important to agree upon the method and frequency of data transmission from the field to the central office and to ensure it is monitored. Depending on the organizational structure, data can be transmitted in real time or at established hours (usually by night) directly to a central server, or may pass through intermediate nodes like regional offices or supervisors. For automatically collected paradata, it is important to ensure it is transmitted when interviewers' devices reach Wi-Fi nodes, for example. Either way, it is important to use secure methods such as encrypted file transfers.

175. Once paradata are collected, programmers need to extract and process them to create KPIs. It is essential to allocate IT resources for data manipulation, storage, and analysis. Additionally, appropriate software for data processing and statistical analysis should be selected, tested, and refined (O'Hare and Jans, 2012).

176. ***Example*:** The Survey of Health, Ageing and Retirement in Europe (SHARE) uses an electronic sample management system (SMS) called the "Case Management System" (CMS) developed by

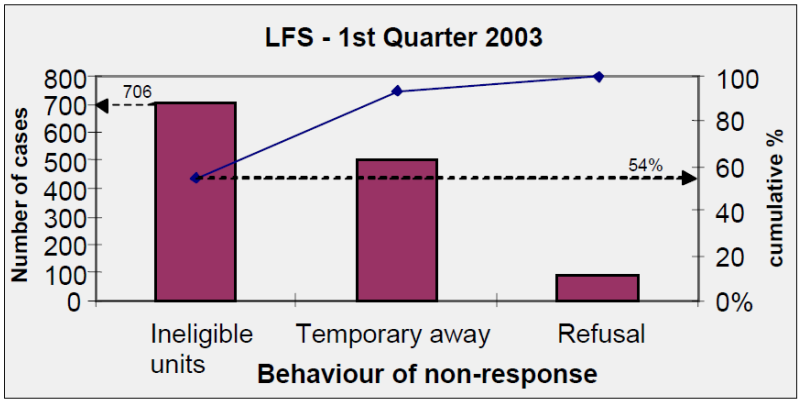
CentERdata, which monitors the survey progress in real-time. Bi-weekly reports of project-specific indicators are generated for the SHARE coordinating team.

6.7.3.2. Step #2: Visualizing indicators

177. Visualization techniques range from more simple tools such as histograms, Pareto charts, scatterplots, and contingency tables to more sophisticated instruments such as cause-and-effect diagrams and control charts (Kreuter et al., 2010). Below we discuss several of these options in more detail.

178. The **Pareto chart** is a graphical tool that displays the contributions of different sources or groups to a total effect or error. The Pareto chart features two components: a bar chart displaying the causes in order of frequency, from most to least, and a line graph showing the cumulative percentage of contributions from each factor (Tagaram and Chen, 2024). Pareto charts allow an analyst to identify the subset of factors that have the greatest impact and/or greatest opportunity for improvement. For example, in the context of the Labour Force Survey (LFS) in Portugal, a Pareto chart displaying different types of non-response showed that "lost" (ineligible) units accounted for 54% of non-responses (**Figure 6.1**). This insight highlights the importance of investigating high non-responses rate to identify underlying causes and take corrective actions (Jones et al., 2003; Bergdahl et al., 2007).

Figure 6.1. Example of a Pareto Chart.



179. A **contingency table** displays the frequency distribution of categorical variables (e.g., performance indicators by interviewer, week, geographical area, etc.). To avoid an overwhelming amount of data, they are best utilized for a focused set of variables and cases. A more streamlined approach is to use **control** charts. These charts plot a selected measure(s) (vertical axis) across time (horizontal axis), with the goal of detecting out-of-ordinary outcomes (Fazzi et al., 2022; Kreuter et al., 2010). A predominant type is the Shewhart p-chart. In this case, the central line tracks the performance measure across time and lines for upper and lower limits demarcate the acceptable range of variation. Typically, the upper and lower limits are established at three standard deviations from the center line, which captures the process average (Statistics Canada, 2003). In survey organizations, Shewhart p-charts are frequently employed to evaluate nonresponse reduction strategies throughout the data collection period (Survey Research Center, 2010). An example of a Shewhart p-chart is shown in **Figure 6.2**.

180. The **cause-and-effect diagram**, also known as the fishbone or Ishikawa diagram, is a tool used to visualize the root causes of a problem. In the diagram, the head of the fish, typically on the right side, represents the problem or area needing improvement, while each branch corresponds to a major category

of causes related to the effect. The diagram allows a team to display possible causes in increasing detail, making it useful for brainstorming potential solutions. Typically, the most significant factors (often five or six) are selected for measurement and improvement (Tagaram and Chen, 2024; European Commission Handbook, 2007). An example of a cause-and-effect diagram is shown in **Figure 6.3**.

Figure 6.2. Example of a Shewhart P-Chart (From Jans et al., 2013).

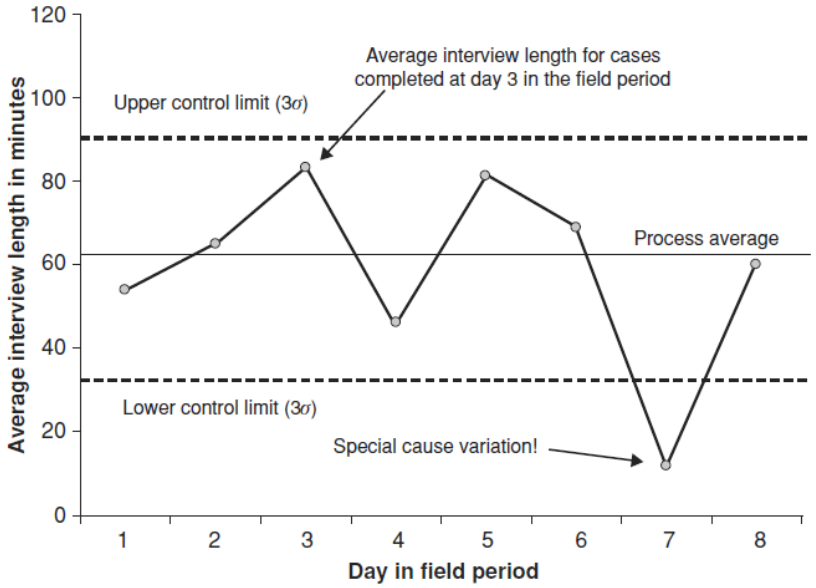
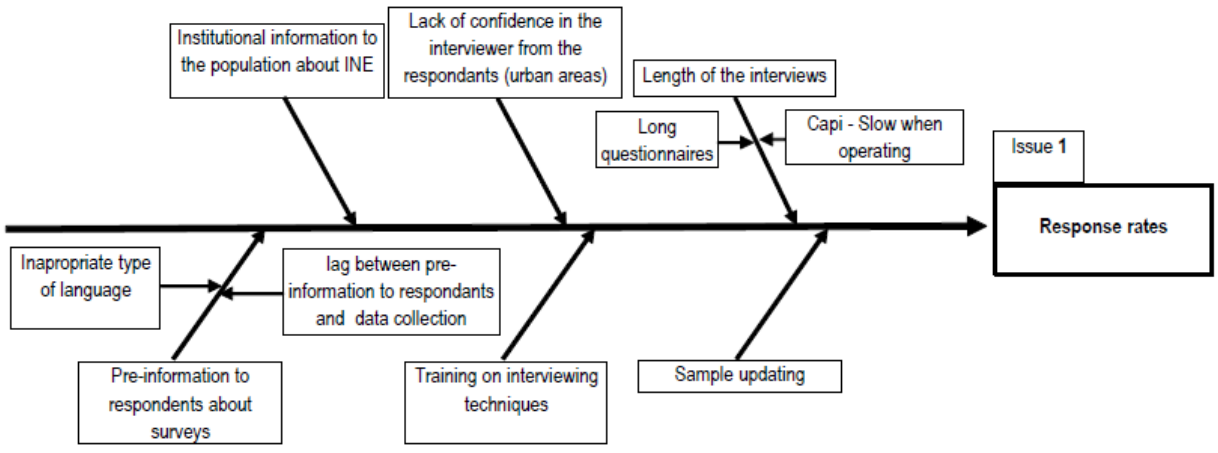


Figure 6.3. Example of a Cause-and-Effect Diagram.



181. Multiple tools exist for the creation of data analysis and visualizations. One that has been gaining popularity is Microsoft Power BI – a software platform designed to convert raw data into sophisticated visualizations and engaging reports. Its primary strength lies in its high capacity for multiple types of data visualization outputs, enabling users to create interactive charts, graphs, maps, and tables. These features

facilitate data analysis from various perspectives, making it easy to identify trends, patterns, and anomalies. Power BI Service, the cloud component of Power BI, and its mobile components foster collaboration "on the go" by allowing teams and organizations to share reports and dashboards (Metre et al., 2023). Below are two examples of uses of approaches to data visualization in production monitoring for survey data collection.

182. ***Example:*** The Statistical Office of the Republic of Serbia (SORS) has adopted the Power BI tool to enhance its statistical operations amid the COVID-19 pandemic and challenging working conditions. This tool was introduced to generate various types of reports and to analyze and compare data across different statistical surveys. The interactive Power BI reports are organized into multiple pages, each providing information about the survey's progress and data collection process. By leveraging Power BI for data analysis, users gain immediate access to raw data, which aids in the deduction process and offers a more accurate interpretation of key indicators. This method streamlines data analysis by sectors, activities, and regions; the team tracks these elements on a monthly basis, allowing for a comprehensive and detailed examination of results. The approach has made it easier to identify which activities, sectors, or regions are most affected by the pandemic or other influences (Hinda and Tolić, 2021).

183. ***Example***: Bieber et al. (2020) introduced a visual strategy using the German Longitudinal Election Study (GLES) to improve fieldwork monitoring in face-to-face interviews, by mapping key indicators geographically in Stata. This spatial visualization offers several advantages. It provides a quick, preliminary overview, directing attention to areas of concern and enabling the reallocation of interviewers from high-performing to lower-performing regions. The tool is also highly adaptable. It allows visualization at various levels (e.g., federal state, municipality, interviewer, or respondent addresses) depending on the study's needs. Furthermore, both the indicators and the threshold values (e.g., color coding) can be customized to meet specific study requirements. While geographic tools are particularly effective for managing face-to-face interviews, they can also be adapted to other data collection modes if respondent or interviewer location data is available.

6.7.4. Monitoring

184. As soon as data collection begins, the survey manager(s) begins production monitoring. For example, those overseeing production monitoring should track response rates and provide feedback to improve interviewer performance and address any issues. In face-to-face data collection, some monitoring of operations will be assigned to fieldwork supervisors who manage interviewer assignments, affirm overall targets are met, and ensure any necessary corrections and adjustments (Statistics Canada, 2003). During fieldwork, team members must maintain regular communication to address unexpected challenges in the field, modify the design (in the context of a responsive survey design approach), and provide on-the-job training to interviewers as necessary.

185. An illustration of possible interventions that can be implemented when out-of-bounds outcomes are detected in production monitoring is provided in **Figure 6.4** (Fazzi et al., 2022; note that UCL and LCL refer to the upper and lower control limits). The content of the exemplar chart shows once again that there is some overlap between concerns at the center of quality control efforts (e.g., interviewer fabrication) when production monitoring is implemented, underscoring the need for close coordination between relevant team members before and during data collection.

Figure 6.4. Example of interventions for above or below pre-established limits.

TYPES OF OUT-OF-CONTROL EVENT				POSSIBLE ACTIONS
RESPONSE RATE	ACTIVITY RATE	ELIGIBILITY RATE	NON-INTERVIEW RATE	
Below the LCL or in control	Above the UCL	Below the LCL (too many ineligible units)	Below the LCL or in control	*Interviewer to be checked:* he/she might cheat in surveying ineligible units (they are given a short questionnaire that is paid as a completed interview)
Below the LCL	Above the UCL	--	Above the UCL or in control	*Interviewer to be re-trained or generalized problem:* he/she shows a high activity rate (*i.e.*, he/she assigns a lot of final outcomes), but this is due to a high non-interview rate. It is important to understand if he/she faces any difficulties in the contact phase or if the contact information is not updated. If the entire province is flagged, then the problem of outdated contact information is generalized
Below the LCL	--	Above the UCL	--	*Interviewer to be re-trained*: units surveyed by him/her are mostly eligible, but he/she has problems in completing the interview
Below the LCL	Below the LCL	--	--	*Interviewer to be checked:* he/she is slow and possibly lazy and should be invited to put more effort in his/her work

6.7.5. Dynamic, responsive, and adaptive design

186. The ability to monitor continually the streams of process data and survey data allows for a more flexible approach to data collection, creating opportunities to adjust the design during data collection. This flexibility can improve survey cost efficiency and leads to more precise, less biased estimates (Groves and Heeringa, 2008). Adjusting a survey design based on data obtained from one stage or from a previous similar data collection is referred to as dynamic, responsive, or adaptive design (Tourangeau et al. 2018).

187. Responsive designs initially involved sub-setting a survey into stages, often introducing experimental (e.g., split-sample) approaches into first stages, and then making evidence-based decisions to shift approaches in later stages to increase efficiency (e.g., increase incentives, assign particularly effective interviewers).

188. Adaptive survey designs (ASD), assume that different people or households may receive different treatments; typically these are defined before the survey and, when combined with responsive design approaches, can also be adjusted using data linked to the survey sample or paradata (Calinescu et al., 2012).

189. Both approaches, and myriad variations (e.g., doing away with specific stages and updating as needed rather than at pre-determined intervals), can be considered types of dynamic design that permit data collection efforts to continuously make improvements by integrating paradata and other observations while data collection is in progress (Tourangeau et al. 2018).

190. Below are two examples of dynamic design during production monitoring. More details on the theory and concepts behind adaptive and responsive survey designs, as well as examples based on virtual and on real survey data, can be found in a large body of published work on these topics, including Calinescu et al. (2013), Grooves and Heeringa (2006), Lundquist and Sarndal (2013), Schouten et al. (2013), Tourangeau et al. (2018).

191. ***Example***: The 2018 Nigerian HIV/AIDS Indicator and Impact Survey (NAIIS) adopted a responsive survey design, that used real-time monitoring of key performance indicators to make improvements during data collection. Paradata (questionnaire completion rate, blood draw response rate, and others) were displayed on a dashboard and updated throughout the survey. Monitoring real-time data reporting led to adjustments in laboratory distribution and staff supplies. These data also helped identify and resolve errors or inconsistencies in the data (e.g., more than one address in the sample), which were corrected during the data collection phase. Project leaders estimate that responsive design saved about 700 hours of fieldwork by addressing errors in the field, avoiding the need for teams to return to enumeration areas; they further estimate the total effort saved US$4.4 million, most of which in the form of payments for security personnel, transportation, and staff (Jahun et al., 2018; Carletto et al., 2022).

192. ***Example***: In 2009, the Swedish Living Conditions Survey conducted an experiment to explore various adaptive survey design strategies. Eight key sample subgroups were identified, and data collection was considered complete once the response rate for each subgroup reached a specific threshold. Specifically, during this process, the R-indicator (with R for Representativity; Schouten et al. 2009) was used to assess the balance of respondents based on key characteristics and the difference between respondents and nonrespondents. The results indicated that implementing appropriate interventions in the data collection design could lead to significant improvements, such as increased balance and reduced discrepancies, compared to the standard LCS data collection process. Additionally, the team found that these interventions could lower data collection costs by requiring significantly fewer call attempts (Lundquist and Sarndal, 2013; Carletto et al., 2022).

6.7.6. Report

193. Accurate documentation throughout the data collection process is essential. This step ensures that all records related to planning, monitoring, control, and improvement are preserved (Survey Research Center, 2010). Reports should evaluate organizational performance against standardized indicators, and include lessons learned and improvement recommendations (Iarossi, 2006).

194. Key documents should cover sample management, paradata, field staff structure, and quality control procedures (Survey Research Center, 2010). Comprehensive assessments should detail performance, corrective actions, baseline updates, and revisions to quality plans. Regular monitoring and reporting on resource use, expenditures, and progress is crucial. Team members must follow a standardized reporting process, providing updates on these aspects, along with operational data like response rates and quality control insights. The frequency of reporting should be agreed upon and adjusted based on the urgency of addressing any issues that arise (Statistics Canada, 2003).

195. Effective process documentation allows for timely interventions and, as well, facilitates the creation of a quality profile for the project. As documented by the University of Michigan's Survey Research Center (2010), a quality profile is a report that documents all protocols and methods deployed in the project, provides indicators related to quality relevant to TSE and more (e.g., timeliness), and registers lessons learned and, if possible, makes recommendations for similar studies. The audience for such a report varies, from internal use to dissemination to data users, so that they can more easily analyse and interpret the survey results.

6.8. Quality control

196. The objective of quality control is to maximize adherence to the study protocols and minimize error. This effort requires multi-faceted and ongoing checks. This section first introduces quality control in data collection (hereafter, QC) as one component of an overall quality assurance (QA) process, and next

identifies considerations relevant to QC planning and design. The section then introduces the types of data that can be monitored for QC and discusses approaches to each data type in turn.

197. This section addresses these topics:

- Quality assurance and control (Section 6.8.1)
- Planning and design (Section 6.8.2)
- Paradata review (Section 6.8.3)
- Substantive data review (Section 6.8.4)
- Audio review (Section 6.8.5)
- Visual review (Section 6.8.6)
- Location tracking (Section 6.8.7)
- Verification (Section 6.8.8)

6.8.1. Quality assurance and control

198. Ensuring survey quality is an effort that starts before fieldwork begins. It requires a holistic consideration of all aspects of the survey process, including the design of the survey instrument, recruitment of team members, training of the data collection team, and logistical planning for the survey. *Quality assurance (QA)* involves establishing processes that set up interviewers and field research organizations for success in realizing high quality data collection (Lyberg & Biemer 2012). For example, survey instruments that optimize on clarity, relevance and flow increase respondent engagement and make the interviewer's job easier (see Chapter 4). Effective training (see Section 6.4) minimizes unintentional errors in data collection; in addition, by boosting skills, trust in the process, and pride in the effort, training can increase buy-in and make intentional deviations from standards less likely. These considerations extend further. For example, quality assurance includes ensuring that all individuals involved in designing, implementing, and monitoring the survey are fairly compensated and feel that they have a stake in the research process so as to incentivize high quality work.

199. Of key relevance to this section, a strong quality assurance system requires planning and designing an approach to monitoring – reviewing, auditing, and verifying – adherence to the survey's data collection protocols. *Quality control (QC)* in data collection is the implementation of that system.

200. A robust approach to QC is one that captures and monitors multiple indicators. More constant attention may be placed on some indicators, while data related to other indicators may be reviewed only if an interview or interviewer is flagged for a suspected deviation from protocols. Consider, for example, timing data show that an interview has been conducted at an unusual time of day. The auditing team may suspect a failure to follow protocols; to check, the auditing team then examines other QC indicators to determine whether, in fact, the suspected deviation is evidence of failure to follow protocols (whether unintentional or not) or, instead, a valid outcome.

6.8.2. Planning and design

201. Planning QC for data collection involves three core considerations: organizational culture; resources; and project protocols and standards. *Organizational culture* starts from the top: leadership must be fully committed to data quality. This norm needs to be transmitted into messaging so that data collection team members take pride in and, ideally, are rewarded for high quality outcomes. Effective leadership is also important when anticipating challenges and adopting new methodologies; project leaders should be prepared to help support the field research organization and its members as they navigate these issues.

Supportive actions include having an expert join in the training and the first few days of fieldwork, fielding calls throughout, or leading training sessions to help explain best practices for the new approaches.

202. *Resources* will constrain what is possible within a QC system, but it is important to keep in mind that costs in the short-run can translate into lower costs in the long-run (e.g., replacing poor-quality interviews while data collection is in progress is less expensive than replacing interviews after data collection has been closed). Many risks can be mitigated with small investments. Where trade-offs must be made, it may be prudent to prioritize those elements most at risk of introducing substantial error. For example, if on-the-ground conditions make fieldwork challenging, these conditions could incentivize interviewers to deviate from the sampling protocols; a holistic approach is to provide interviewers with adequate support in the field, while using a QC approach that monitors and verifies that interviewers carry out their work in the locations required to realize the intended sample.

203. *Project protocols and standards* are a critical input to planning a QC system for data collection. Project protocols include sampling locations, respondent selection and recruitment, questionnaire instrument and instructions, the recording of paradata, among other items. QC standards need to be established prior to the beginning of data collection to ensure agreement on how data collection is monitored and verified. This might involve, for example, minimum thresholds for the length of an interview (whether measured by time and/or the percentage of questions asked and answered before the interview ends or breaks-off). It can also involve standard-setting around deviations from the intended location of an interview, expectations regarding how quickly and verbatim the survey questions are asked, and the identification of statistical patterns in the data that suggest fabrication.

204. Once standards are established, the QC plan needs a list of *flags* – that is, thresholds or instances when a survey will be considered potentially deviant. When fieldwork begins, QC requires constant checks on the team's processes and on the data that is being collected. Recent advances in technology, particularly the capture of data using computer tablets or other e-devices, greatly assist in this process by generating efficiencies (e.g., software can be programmed to raise flags, in place of manual review) and by permitting QC to operate in real-time, or near real-time. The collection and auditing of e-data at the time of the interview, or soon after (when, for example, internet access is available), allows for immediate feedback for interviewers who may be collecting lower quality interviews, including the ability to retrain them. Or, if an interviewer is deliberately fabricating responses, the individual can be removed from the survey team. Monitoring results throughout the process allows for replacement interviews to be conducted prior to the completion of data collection; this is cost-saving in the case of face-to-face surveys when such replacements take place before the team moves on to a new geographic location (Cohen and Larrea 2018; Logan et al. 2020).

205. The design of quality control processes will vary by survey mode. For face-to-face surveys, extensive planning and oversight of the field team is required. For CATI surveys, particularly if all the interviewers are in the same call centre using automated dialling, QC processes will typically prioritize listening in on calls for quality assessment. Meanwhile, for web-based interviews, a greater emphasis will fall on ensuring that the person taking the survey is a member of the target population and responds in a serious manner. For web surveys, QC steps that can be taken include repeating questions to ensure consistent responses; asking questions designed to measure if the respondent is paying attention; requiring open-ended comments to analyse response; preventing multiple responses from the same IP address; device fingerprinting to ensure the same device is not used for multiple entries; and validating email addresses (see, e.g., Anduiza and Galais 2017; Goodrich et al. 2023; Pinzón et al. 2024).

206. Regardless of the survey mode, it is necessary to carefully consider how quality control procedures will be implemented. The use of technology for data collection significantly increases the scope of what can be monitored for QC. In a comprehensive study, Cohen and Warner (2018) identify 141 distinct QC measurements that can be collected with the assistance of technology in the data collection process. Not all these items are equally useful; from that larger set, the authors identify 30 QC measurements that are comparatively high-performing. In brief, pre-planning around a smaller set of QC items that are carefully deployed and analysed can be both more efficient and more effective than collecting all possible data. In addition, the use of technology to record data and transfer in the field must be planned in accordance with local rules, local norms, and privacy considerations. **Table 6.2** lists six core types of information that can be monitored, audited, and/or verified during the data collection process. The table provides a summary definition and statement on use. The sections that follow discuss each type in more detail.

Table 6.2. QC in data collection: Types of data and use.

Data type	Definition	Types of use
Paradata	Data collected about interviews and the survey process	Assessing unexpected implementation outcomes such as short interview duration, unlikely hour of the day, or irregularities in response rates for interviewers
Substantive	Core information (typically, responses) gathered in interviews	Examining response patterns for illogical consistency, deviant statistical patterns, item non-response, or satisficing
Audio	Recordings of oral components in interviews	Review to determine if questions were read accurately and at a reasonable pace, responses were entered accurately, or if items were not administered
Visual	Photographs of aspects related to the interview	Review of photos of the (general) location of the interview to compare with GPS data, photos of the interviewer to verify their identity or establish location, or (rare, given privacy safeguards and regulations) photos of the respondent to ensure different respondents are taking the survey
GIS	Geolocation coordinates recorded by the software program in face-to-face interviews	Ensuring the interview is occurring in the intended location
Verification	Third party checking of interviews	Supervisor observation of the interview or back-checking responses by a supervisor or other third party who re-administers part of the survey to the intended respondent to ensure the answers are consistent with those recorded by the interviewer

6.8.3. Paradata review

207. Paradata contain data about the process of data collection. Paradata review is a central component of production monitoring, as discussed earlier, and it also facilitates quality control auditing. Paradata relevant to quality control may include information on the days and times of interviews, the number of contact attempts, number of partial and completed interviews, and response rates by interviewers. When data are collected passively using technology and not from the interviewer's self-reporting, then they allow for independent insight into the interview itself. Even when technology is used, some paradata will be reported by interviewers (e.g., contact attempts), whether using the e-device or not. When technology is not available, interviewers will need to record all paradata either during or soon after the particular data collection effort. Some projects collect paradata via self-administered surveys that are completed by the interviewer within a certain pre-determined time frame. When interviewers are self-reporting paradata, some portion of training and QC effort should be allocated to ensuring this is done in an efficient and accurate manner.

208. Timing data permit a review of the overall duration of the interview. These data can reveal if the interview was rushed or took an extremely long time. When data are captured using e-devices, timers can be programmed throughout the survey to check if particular sections are rushed (or possibly skipped entirely) or to detect breaks during the interview. Such information can point to issues that need to be addressed with the enumerator or may point to issues that need to be addressed within the questionnaire – for example, if there are certain segments that consistently run long that could signal a lack of understanding by respondents.

209. In some cases, timing data provide information about the quality of the sample. For example, for in-person surveys that require moving from one house to another, interviewers need to identify the next house per the sampling plan and most likely they will encounter refusals or the selected respondent may not be home. If paradata show there is an unreasonably short time between interviews, it may be that interviews are valid, but that respondent selection is deviating from the sampling plan.

210. Other aspects of paradata are self-reported by the interviewer. For example, the number of contact attempts and the result of each one give insight into the selection process. If an interviewer never records a refusal, then this would be surprising; most likely, the interviewer is not accurately recording attempts or not implementing the sampling plan. Or, if an interviewer has an excessively high number of refusals compared with others, it may be beneficial to speak with them or consider retraining them on how to engage a potential respondent upon first contact (see Section 6.6 of this chapter). For CAPI or PAPI surveys, there may be geographic clustering of some of these aspects, meaning that in particularly difficult areas, refusal rates will be higher, so it is necessary to consider whether variations are interviewer- or respondent-driven. This often can be done by comparing interviewers on the same team who conduct interviews in the same locations or otherwise similar situations (e.g., Goel and Abraham 2024).

211. In a study of interviewer performance in the India Working Study, Goel and Abraham (2024) identify three paradata indicators that were particularly effective in improving performance: number of completed interviews; average interview time; and ratio of completed to started interviews. They use a process of "dynamic benchmarking", in which they compared interviewer performance (on a weekly basis) against a moving average and flagged deviations that exceeded a pre-established threshold. In their case, they used both 1 and 1.6 standard deviations from the mean; while selecting an initial threshold is arbitrary, the decision can be updated following pilot studies and/or as QC for data collection unfolds. Goel and Abraham (2024) report that when the process identified actionable interventions (e.g., a private coaching conversation

between the supervisor and interviewer), those were undertaken in a timely manner in such a way that resulted in quantifiable improvements in data quality.

212. Other paradata that can be recorded about the interview include the following: whether there was a bystander present in addition to the interviewer and the respondent; an assessment by the interviewer about how the respondent answered questions - quickly, reluctantly, or seemingly distracted, for example; perceptions of whether the respondent fully understood the questions; or respondent feedback about the interview including any complaints about it.

213. Additional types of paradata for in person surveys include geographic locations, which will be discussed separately. For phone surveys, particularly if not autodialled, then the phone numbers themselves serve as a form of paradata that can be assessed to make sure interviewers are calling the numbers that are given to them from the sample.

214. For web-administered surveys, paradata will differ and most will relate to the interviewee (e.g., from where and when the survey is submitted, the speed at which the respondent completes parts or all of the survey, and other data that facilitate assessments of the integrity of the respondent and the quality of their responses). In brief, for web surveys and other surveys conducted without interviewers, the main focus of paradata in QC processes is to ensure that the correct individual is filling out the survey and that this person is doing so in a thoughtful way that represents her or his true attitudes.

6.8.4. Substantive data review

215. Throughout data collection, it is important to review the response data for a number of elements. This includes checks on item nonresponse, logic checks (e.g., for expected correlations), and unexpected outcomes (e.g., means that are atypical for the population). In addition, more detailed approaches include analyses that test for interviewer effects and/or statistical packages that help identify questionable patterns in the data.

6.8.4.1. Item nonresponse

216. Assessing item non-response is critical to identifying weaknesses within the questionnaire. Item non-response may indicate that respondents are not understanding a question, that they are uncomfortable or scared to answer it, or that interviewers are uncomfortable asking it and are recording “don’t know” or “refuse” without asking the question to the respondent.

6.8.4.2. Logic checks

217. Logic checks are analyses that focus on expected associations between different items on the survey. For example, a person who is not a smoker is unlikely to report having purchased cigarettes in the past week. Of course, it is possible that the respondent is buying them for a family member who does smoke. Thus, like other data collected in QC processes, individual data points are typically suggestive and not definitive of quality issues; for this reason, QC in data collection functions best when multiple indicators are assessed. In addition to looking for logical inconsistencies, it is possible to test for expected correlations; for example, if there is no correlation between smokers and buying cigarettes in the past week, then it would suggest a problem that needs to be investigated. The first steps in such an investigation are to determine if the instrument was written and/or programmed correctly and that the data processing is valid; if so, then such an unexpected pattern might indicate some type of fabrication or other major data quality issue introduced during the data collection process.

6.8.4.3. Unexpected or implausible outcomes

218. Beyond direct logic checks, it is important to examine the data based on knowledge about the country or target population. If the team has access to similar questions asked on similar surveys in the past, these can also provide perspective on what constitutes a deviant outcome. Survey results can be surprising, but if results are contrary to most expectations, then further research may be necessary to understand the reason for the results or to ensure that it is not measurement error or some other form of error introduced during fieldwork. Sometimes valid results may at first appear quite unusual. The highest rating in the history of polling of a U.S. president occurred directly after the September 11 terrorist attacks. Instead of losing faith in the president for failing to stop the attacks, the United States public drastically increased their support for the president due to a "rally-around-the-flag" effect (Kam and Ramos 2008). Likewise, if a major natural disaster struck a region soon before a survey, concern about climate change might show a significant shift from a pre-disaster mean; yet in the absence of such a contextual explanation, a dramatic change in this attitude should be flagged as a QC issue that might be related to some aspect of the sample, question order, programming or processing problem, or other factor.

6.8.4.4. Interviewer effects

219. Examining survey data can also provide insight into how aspects of the data collection implementation process are shaping the quality of the data. One example is conducting analyses of substantive responses against interviewer traits (e.g., gender) to assess whether interviewer effects are biasing responses. For example, if response patterns to questions on gender equality vary, all else equal, conditional on whether the interviewer is a man or a woman, project leaders may need to consider the fieldwork plan, interviewer assignments, or other adjustments.

6.8.4.5. Statistical analysis

220. Substantive review can be facilitated by programs that detect statistical patterns in the data that would not be obvious at first glance (see Robbins 2019). Several established tests permit QC assessments. These include PercentMatch, which seeks to determine whether the response patterns between two distinct interviews are overly similar (Kuriakose and Robbins, 2016); Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) analysis of specific batteries to determine if interviews do not display expected variation within a battery (Blasius and Thiessen 2012 & 2015); and applications of Benford's law for questions that require numerical entry (Swanson et al. 2003). These tests can point to low quality data that could be the result of fabrication.

221. In deciding on a plan for substantive data review, it is important to keep in mind that the types of tests that are most appropriate will vary by type (e.g., mode) and the risks (i.e., the relevant threats to quality) inherent to each survey. Further, as noted above, it is advisable have a plan for triangulating (cross-checking) across different types of QC checks so that, when a flag is raised, it can be further investigated.

6.8.5. Audio review

222. When technology is available it can be possible to record parts of the interview or, capacity-permitting, the entire interview. This approach is generally referred to as Computer Audio-Recorded Interviewing (CARI). CARI should only be used when permitted by local rules, in ways that maintain data privacy (e.g., recordings are kept in a secured location), and with the respondent's informed consent. Most software for CAPI and CATI data collection include recording as a programming option, with voice files uploaded to a secure cloud account on connection with the internet. Having an auditor(s) listen to these recordings can detect issues that may not be apparent from other sources, including not reading the consent form accurately, skipping a question, not reading questions accurately or at an appropriate pace, not entering

responses correctly, or not following instructions such as not reading certain response items or paraphrasing items in the questionnaire. Audio files can also be checked to ensure there is a respondent present.

223. The precise approach will vary by project. If there are data storage or upload constraints, then project leaders should identify items that are particularly important and/or at risk, and prioritize these for recording. In addition to informing respondents that audio is being recorded for quality control purposes, ethics around transparency require letting interviewers know their work is being monitored; yet, when only parts of the interview are being recorded, it is best not to disclose which questions are being audited. In other cases, it is possible to record the entirety of each interview. A benefit is that these recordings can be retrieved as part of a QC investigation; for example, if timing data show an interview is suspiciously short, or an interview fails certain substantive logic checks, it is then possible to listen to the full interview or to specific sections to ensure that the questions were administered accurately and the responses recorded correctly. Or if a flag is raised on the quality of an interviewer's performance in a handful of cases, it is possible to use this archive to audit all the other interviews by that individual.

224. The percent of an interviewers' work that is audited will vary across projects and, as well, across the duration of a project. For example, a project might begin by auditing a selected set of items for 100% of interviews. Then, as feedback from the QC process boosts average interviewer quality, it may be possible to decrease the amount of auditing of consistently high performing interviewers while maintaining or increasing auditing for lower performing interviewers. Such a trajectory in terms of increases in quality over the course of a survey and distinction between "A" (high performing") and "B" (lower performing) groups of interviewers is documented by Cohen and Larrea (2018). In assessing QC items, Cohen and Warner (2021) find that a set of audio review indicators are most determinative of whether an interview is canceled for quality reasons: failing to read part or all of the consent form; skipping one or more questions; interpreting or misreading questions; and no respondent detected in the audio review.

225. Interviewers will inevitably make some errors (e.g., misreading a word or phrase) in administering surveys, particularly long ones. Interviewer abilities vary and some have voices or accents that are easier or harder to understand for a respondent, while some naturally speak faster or slower. The quality of the interview depends, at least in part, on the quality of both the interviewer and the respondent and how they interact with one another. In brief, it is reasonable to expect and accept some variation in quality across interviews and across interviewers; performing audio review permits a deeper understanding of the prevalence of minor versus major deviations from protocols.

226. Audio reviews can be used to identify challenges to data quality that extend beyond the interviewer. For example, if the respondent asks the interviewer to repeat the question numerous times, to slow down, or says they do not understand, this suggests a comprehension issue. If that issue is persistent across interviews, it may be that the questionnaire item(s) are not sufficiently well-constructed (see Chapter 4 on best practices in questionnaire design).

227. As with most QC processes, it is essential to pre-establish standardized indicators for audio review. If there are certain questions that are essential to be completed correctly, these should be noted. If an interviewer accidentally misreads a critical part of a survey instrument, it may be necessary remove the entire interview from the final dataset. Alternatively, if the interviewer were to misread the question on the respondent's age, but then correct the language, then this error would be expected to have minimal effect. When using CARI, the results from the analysis should be clearly summarized in reports with clear reasoning given for each decision. These can then be tracked over time to ensure that improvements are being made in the quality of interviews and adherence to protocols.

228. The AmericasBarometer – a biennial survey of Western Hemisphere countries directed by LAPOP Lab at Vanderbilt University – regularly uses audio recordings to coach better performance out of interviewers (e.g., tailored feedback to interviewers who read too fast or full retrainings) and, as well, to recognize excellent work. In that project, the study's leaders pre-determine numerical penalties for a range of potential deviations from protocols – including for a range of quality control errors. More problematic errors (e.g., not reading the consent form) receive higher penalties. If an interview's audit results in a value above a pre-established threshold, the interview is cancelled and replaced while data collection is in-progress (Cohen and Larrea 2018; Cohen and Warner 2021).

6.8.6. Visual review

229. For CAPI surveys, it is possible to use the computer tablet to capture images that can help with quality control. One approach is to capture an image of the interviewer as she or he is administering the survey. This can be used as a check on *who* is conducting the interview and *where* the interview is taking place. Certain software programs for conducting in-person surveys using tablets permit code that takes a forward-facing (that is, facing the interviewer and not the interviewee) picture at a pre-selected time during the interview. These image files, which can be loaded to a secure cloud location on connection to the internet, can be used to ensure that the interviewer has conducted the interview and not outsourced it to another individual who has not been trained and certified to carry out the survey project.

230. For face-to-face projects in which each interview is located in a different place, images should vary across interviews. With advances in technology, collecting image data is not resource-intensive, yet because reviewing image files is labour-intensive, these data might only be consulted if other flags raise concerns about an interviewer's performance.

231. In still other cases, image files might be used in in-person panel studies, where it is important to verify that the same location was reached across waves. For example, interviewers can be instructed to take pictures of the general entrance to the selected dwelling to match up dwellings across the first and subsequent contact attempts.

232. Resources and risk assessments matter in determining whether image files are reviewed for 100% of interviews, spot-checked, or only audited when other flags lead the QC team to dive more deeply into an interviewer's digital record. Manual reviews of image files, especially when they require cross-checking the image against other images or other data sources, are by nature time consuming and therefore typically not recommended as a primary approach to QC in data collection. However, if there are other questions about interview quality or if the images are needed for other purposes (e.g., verifying return to the correct dwelling in a panel study), then review of image files becomes a useful option.

233. One caution is that image files may reveal information about the respondent or provide identifying information about the respondent's place of residence. It is important to train interviewers to avoid capturing identifying images unless permitted as part of the protocol. Further, it is critical that these photos be stored securely and after verification purposes, not linked to the specific interview. It is typically mandatory to delete image files after a project is complete. As with audio files, image files should not be collected if doing so would violate local rules or norms.

234. It may be possible for web-administered surveys to include some form of photo identification if the respondent is part of a non-probability sample such as a pre-existing panel. Requiring the respondent to take his or her picture on whatever device is being used for the survey can help ensure that the intended respondent is the one completing the survey. This image could be compared to one that was taken when the person signed up for the panel or might be used as a basic check against responses to demographic

questions. If the demographics indicate a 70-year-old-female but the image presents as a young male, it suggests the respondent is not accurately reporting key demographic information.

6.8.7. Location tracking

235. For CAPI surveys, it is possible to leverage GPS to assist with QC in data collection. This can be done by using technology to collect GPS coordinates for each interview, program geo-fences into software that can flag out-of-bounds interviews, and collect GPS coordinates for interviewers' routes (similar to how fitness devices map running routes).

236. Prior to discussing each of these uses, three caveats are important to issue around geolocation data. First, for various reasons, geolocation data is not always accurate (e.g., because of factors that obstruct the transmission of that data such as foliage and buildings). Second, and related, it is important to ensure that the collection of geolocation data is not obstructed by setting electronic devices to "airplane mode" or by use of software that spoofs the device's geolocation information. Third, the collection of location data for interviews carries an inherent risk of breach of data confidentiality. Geolocation data should not be collected when it is not legally permissible to do so (i.e., when data privacy laws prohibit it) nor when the unintentional disclosure of those data could put individuals at significant risk (e.g., surveys in conflict zones that ask about support for rebel groups).

6.8.7.1. GPS coordinates

237. A straightforward approach is to collect the GPS coordinates (longitude, latitude) of interviews. These data can be monitored for signs of deviation from the sample design or protocols. For example, if interview coordinates for a household survey in rural area are all tightly clustered, this suggests that interviewers did not reach the intended locations. GPS coordinates can be placed into Google Earth or other software to check whether an interview is in the correct place; since manually checking of GPS coordinates is labour-intensive, this might be used only when other flags suggest a QC issue that warrants further investigation.

6.8.7.2. Geofences

238. If sufficient GIS files are available or can be developed (i.e., shapefiles that capture location boundaries), it is possible to geo-fence sample locations. Software for survey research is increasingly adding this to the suite of programming options; for example, SurveyToGo has this as a built-in capacity. Geofences are boundaries around a sample location. By using software to compare GPS data to geofence data, one can ensure that the interviewer is within the intended PSU or the location specified in the sample (Montalvo et al. 2018). Certain software for data collection (e.g., SurveyToGo) allows the option of programming a flag that is raised for the interviewer on their device (e.g., asking them to check their location) and recording that flag for the QC team to review.

6.8.7.3. Interviewer route tracking

239. Some software (e.g., SurveyToGo) can be programmed to track the route taken by the interviewers to each location. If their location remains constant, that might indicate a deviation from protocols (unless interviews are taking place, for example, in a single apartment building). While software advances permit these data to be collected and uploaded to the cloud for QC use, it can be labour-intensive to review each interviewers' route. Therefore, these data might be best used for spot-checking that protocols such as random walk designs are being followed (e.g., by reviewing a random subset of routes) or in cases where other flags warrant a deeper investigation into an interviewers' behaviour in the field.

6.8.8. Verification

240. A long-standing practice in survey research is a "back-check" verification of the interview by a supervisor or other auditor. Back checking can reveal a number of QC issues. First, it can be used to identify fabrication. If a respondent says that no interviewer visited the house, for example, it could indicate deviation (intentional or unintentional) from the sampling plan. Similarly, if the back check reveals that many of the answers from the respondent differ from those recorded by the interviewer, it could signal fabrication. It is important to recognize that it is not unusual for respondents to provide slightly different responses during the back check (e.g., moving from "agree" to "strongly agree" on a Likert scale might simply indicate the person was ambivalent between those options).

241. Additionally, back checks can also reveal helpful data on the quality of the interviewer, including by asking the respondent if the interviewer provided a clear identification, the purpose of the survey, spoke clearly, and was polite. Each of these data points can be used as feedback to help improve interviewer performance for future interviews.

242. In face-to-face interviews, two approaches are common for back checks. First, the supervisor traveling with the team will go to the house and verify the interview, including asking if a person in the household completed an interview and then repeating a small subset of the questionnaire to ensure that the answers (mostly) match those recorded in the interview. A second approach is a callback by phone assuming that the respondent provided a telephone number. Typically, this is done by the central office in the days following the interview and repeats the process. Both practices can be outsourced to an external firm not connected with the original firm as a further form of verification; this approach tends to be more reliable in the case of callbacks because there is no conflict of interest.

243. Callbacks can also be used for QC in CATI interviews. However, this approach may only be required if the original interview was not recorded allowing for verification in this manner. For web-administered surveys, back-checks tend to be significantly more complicated using these standard approaches.

244. When PAPI used for face-to-face surveys, it is common to require random back checks for 25–30% of interviews. As Struwig and Roberts (2020) detail in their description of protocols for the South African Social Attitude Survey (SARAS), it can be optimal to use a mixed approach; in the SARAS project, team members perform in-person checks in real-time during data collection, with an emphasis on the initial stage, to assess and nudge adherence to protocols by both interviewers and supervisors; then, both telephone and in-person back checks are performed at the completion of data collection to validate surveys and solicit feedback.

245. With the rise of CAPI and other advances described that permit extensive audits using paradata, often back checking is only done for a small set of interviews, such as 10%. Teams may decide to vary the percentage of backchecks across interviewers; for example, if QC audits identify consistent problems for an interviewer, it is advisable to conduct additional back checks for that individual's interviews. Overall, the specific percent of interviews to back check requires an assessment of resources and risks, alongside an assessment of the extent to which other QC measurements substitute for backchecks (e.g., robust use of audio, visual, and GIS audits may make most backchecks redundant).

6.9. Emerging approaches

246. Technological advancements have been reshaping all aspects of data collection, from survey mode to training materials to quality control. This section discusses some ways that technology is or can be deployed to create increase efficiency and quality in data collection. This section also anticipates some challenges (see **Box 6.3**) regarding the adoption of new technologies for data collection.

Box 6.3. Potential challenges of new technologies for data collection

When considering new technology and other emerging approaches, it is important to anticipate and address potential hurdles to the adoption of technology, especially in low- and middle-income countries.

- LAWS REGARDING WHAT DATA CAN BE CAPTURED AND HOW. Legal restrictions related to data privacy vary across contexts and across time. Before collecting new types of data, determine local rules on what data can be loaded to the cloud (or stored locally).
- LACK OF CAPACITY OR DISCOMFORT AMONG THE DATA COLLECTION TEAM. When personnel lack skills and/or comfort with new technologies, it can be useful to bring in an outside expert to conduct hands-on training.
- PUBLIC RECEPTIVITY TO TECHNOLOGY MAY SHAPE WILLINGNESS TO COOPERATE. Pilot tests can be used to determine whether new approaches will affect survey outcomes. For example, in a study in Tunisia, Bush and Prather (2019) find switching from PAPI to CAPI did not shift responses to questions about the ruling party.
- SAFETY RISKS TO INTERVIEWERS. In high-crime contexts, low-cost materials can be used to disguise mobile devices (e.g., containers for tablets that look like clipboards), yet theft happens. Therefore, in deploying CAPI, project leaders need to ensure use of a software that encrypts data to prevent third parties from accessing data from stolen devices. In addition, organizational capacity and culture need to be addressed.

6.9.1. Emerging approaches and survey mode

247. Technological advances have created new options for survey mode. Emerging modes include interactive voice response (IVR), short message services (SMS), app-based interviews, and computer-assisted video interviewing (CAVI) (also referred to as Video Mediated Interviewing [VMI] [see Rossetti et al. 2024]).

6.9.1.1. IVR

248. Because IVR is based on pre-set recordings of questionnaire items, it can increase standardization, easily accommodate multi-language interviewing, and may, like self-administered surveys, reduce incentives to censor socially undesirable responses (Firchow and Mac Ginty 2017). Measurement error is likely to decrease as these benefits are realized.

6.9.1.2. SMS

249. As ownership of mobile phones has increased, it likewise becomes easier to collect data via self-administered surveys that are sent via short message services (SMS, aka text messages) or via mobile apps (e.g., Bexelius et al. 2009; Lau et al. 2019; Jäckle et al. 2022); these approaches may be best for very short surveys and/or to circulate information on how to participate in a survey. It is important to consider what effects less personable approaches to data collection may have on satisficing and response bias, while recognizing that such approaches may be particularly useful in surveys on sensitive topics to the extent that they elicit fewer social desirability effects.

6.9.1.3. CAVI

250. CAVI (aka VMI) approaches, which retain a live interviewer who administers the survey via a remote connection (typically a videoconference program), increased during the COVID-19 pandemic. For

example, details on the introduction of this approach into the 2020 American National Election (ANES) Study are provided in a publicly available methodology report (DeBell et al. 2022; see also Centeno et al. 2024). As another example, ISTAT researchers present an analysis of differences between CAPI and VMI approaches in 2022 in the Italian Labour Force Survey (ILFS) (Rossetti et al. 2024).

6.9.1.4. Mixed Mode

251. Using technology to assist in data collection does not need to be a single or static choice. Increasingly, survey projects use any of an increasing variety of mixed mode approaches. In some cases, mixed modes pre-assign potential respondents to one mode or another, while in other cases modes are assigned in a sequential way. For example, in the ANES 2020 study, in-person interviews were prioritized and it was only when potential respondents declined that option that they were offered a CAVI interview. Note that there are different ways to approach sequencing modes: for example, one could prioritize maximizing responses by beginning with more costly approaches that are deemed highly effective in securing responses, and then use less costly approaches in a second phase to try to recruit initial nonresponders; or, one could prioritize cost minimalization and begin with less costly approaches, and then use more costly approaches only in a second phase to recruit initial non-responders. In brief, sequential approaches to mixed mode studies require advance planning in training a mixed method team, determining who will be deployed for each mode and when, and developing a system for documenting and assessing the approach.

6.9.2. Machine learning and other software applications

252. Statisticians use machine learning (ML) to produce computationally complex prediction models. One subset is the large-language models (LLMs) that have spawned generative AI programs such as GPT-4 by OpenAI. The widespread availability of generative AI platforms open opportunities for the development of project-specific chatbots that can answer questions from interviewers (as a complement to fieldwork manuals, for example) and the public (as a complement to publicity activities) as well as chatbots that can facilitate interviews in ways that may increase respondent engagement and, related, data quality (see, e.g., Xiao et al. 2020). As technology advances, so too do opportunities to integrate paradata into such approaches, tailoring them in real time (e.g., prioritizing techniques that are successful in increasing engagement or, in the case of a computer-led interview, adjusting the speed of the interview to maximize attentiveness and data quality).

253. Machine learning approaches can also be used to analyze production and quality control data to find efficiencies. Cohen and Warner (2018) apply a machine learning approach to 141 QC measurements collected as part of LAPOP's AmericasBarometer survey; their analysis yields a subset of comparatively high-performing items that can be prioritized in similar data collection efforts. Likewise, Shah et al. (2020 describe a machine learning approach for a survey in India that runs simultaneous to data collection so that potentially problematic data can be identified in real time. The potential applications of ML-based approaches to review QC data are not limited to data-as-text but, rather, may be applied to images, audio files, and more.

254. Innovations in software are ever-expanding. For example, researchers affiliated with the European Social Survey (ESSE) developed a "virtual surrounding impression" (VSI) tool (Doušak et al. 2024). Running on software that is publicly available (see *Resources*), this software collects location-relevant data using detectable Wi-Fi access points and can be used to flag instances in which more than one CAPI interview appears to be conducted in the same location and/or whether the interviewer changes locations during the interview. Although the feasibility of the VSI approach may be variable across contexts (e.g., conditional on the pervasiveness of Wi-Fi nodes), it has the benefit of not requiring images or GIS location, meaning it maximizes on privacy protection.

6.10. Resources [[INCOMPLETE]]

6.10.1. General guidance

Survey Research Center, Institute for Social Research. 2011. *Guidelines for Best Practice in Cross-Cultural Surveys: Full Guidelines (*3rd Edition). University of Michigan.

https://ccsg.isr.umich.edu/wp-content/uploads/2020/02/CCSG_Guidelines_Archive_2010_Version.pdf

European Social Survey European Research Infrastructure (ESS ERIC) (2024). ESS methodology. Methodology Overview | European Social Survey

6.10.2. Interviewer training

For the specific details on how to plan schedule, and conduct training, see these resources:
Kutka et al. (2023), "A Practical Guide to Fieldwork Training". LSMS Guidebook Washington, D.C.: World Bank Group. https://lsms-worldbank.github.io/pg2sq-training.

SRC 2011, par. X, p.12-15;

Stiegler and Biediger 2016:
https://www.gesis.org/fileadmin/admin/Dateikatalog/pdf/guidelines/interviewer_skills_training_stiegler_biedinger_2016.pdf

6.10.3. Production monitoring

[[NOTE: Need to add links to examples of dashboards here]]

6.10.4. Quality control

PercentMatch:
https://ideas.repec.org/c/boc/bocode/s457984.html#:~:text=Similar%20to%20duplicates%2C%20percentmatch%20compares,match%20percentage%20for%20each%20observation.

See, e.g., Data Collection | European Social Survey

6.10.5. Emerging approaches

For information on an approach that configures tablets for self-administration of a survey, see:
https://www.europeansocialsurvey.org/methodology/methodological-research/modes-data-collection/electronic-questionnaire-device

For details on the Virtual Surrounding Impression (VSI) tool, see this reference:
Doušak, M., Briceno-Rosas, R., & Kappelhof, J. (2024). Virtual surrounding impression: ethical and privacy-respecting tracking of face-to-face computer-assisted personal interview location. Teorija in Praksa, 61(3), 669-686. DOI: 10.51936/tip.61.3.669

And, code for the VSI is available here: https://code.may.si/ may/VirtualSurroundingImpression.

6.11. References

American Association for Public Opinion Research (AAPOR) (2023), 10th Edition of the Standard Definitions report. Retrieved from: Standards-Definitions-10th-edition.pdf (aapor.org), last accessed 5th of August 2024.

Anduiza, Eva, and Carol Galais. "Answering without reading: IMCs and strong satisficing in online surveys." *International Journal of Public Opinion Research* 29, no. 3 (2017): 497-519.

Balì N., Ceccarelli C., Fiori M. T., Lugli A., Rossetti F. 2023a. “Impact of digital learning on the interviewer’s performance”. Rivista Italiana di Economia Demografia e Statistica, 77(1): 65-76. - a

Balì N., Fazzi G., Rossetti F. 2023b. “Efficiency of Surveyors’ Training: In-Person Meetings Versus Distance Learning”. Rivista Italiana di Economia Demografia e Statistica, 77(2): 200-210.

Banks, Randy, and Heather Laurie. "From PAPI to CAPI: the case of the British Household Panel Survey." *Social Science Computer Review* 18, no. 4 (2000): 397-406.

Battaglia, Michael P., Michael W. Link, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad. 2008. "An evaluation of respondent selection methods for household mail surveys." *Public Opinion Quarterly* 72 (3): 459-469.

Bauer, Johannes J. Biases in Random Route Surveys, *Journal of Survey Statistics and Methodology*, Volume 4, Issue 2, June 2016, Pages 263–287, https://doi.org/10.1093/jssam/smw012

Bergdahl, M., Ehling, M., Elvers, E., Földesi, E., Körner, T., Kron, A., Lohauss, P., Mag, K., Morais, V., Nimmergut, A., Sæbø, H., & Timm, U. (2007). Handbook on Data Quality Assessment Methods and Tools. Eurostat Technical Report.

Beullens, K., Loosveldt G., Vandenplas C. & Stoop I. (2018). Response Rates in the European Social Survey: Increasing, Decreasing, or a Matter of Fieldwork Efforts? Survey Methods: Insights from the Field.

Bexelius, Christin, Hanna Merk, Sven Sandin, Alexandra Ekman, Olof Nyrén, Sharon Kühlmann-Berenzon, Annika Linde, and Jan-Eric Litton. "SMS versus telephone interviews for epidemiological data collection: feasibility study estimating influenza vaccination coverage in the Swedish population." *European journal of epidemiology* 24 (2009): 73-81.

Bieber, I., Blumenberg, J. N., Blumenberg, M. S., & Blohm, M. (2020). Using Geospatial Data to Monitor and Optimize Face-to-Face Fieldwork. Survey Methods: Insights from the Field, 1-16 https://doi.org/10.13094/SMIF-2020-00005

Biemer, P. (2010) Total Survey Error: Design, Implementation, and Evaluation, *Public Opinion Quarterly*, Volume 74, Issue 5, Pages 817–848, https://doi.org/10.1093/poq/nfq058

Billiet, Jacques and Loosveldt, Geert. 1988. “Improvement of the quality of responses to factual survey questions by interviewer training”. Public Opinion Quarterly, 52(2): 190-211. https://doi.org/10.1086/269094.

Bischoping, Katherine, and Howard Schuman. "Pens and polls in Nicaragua: An analysis of the 1990 preelection surveys." *American Journal of Political Science* (1992): 331-350.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*(4), 608- 628. https://doi.org/10.1207/S15328007SEM0704_5

Blaydes, Lisa, and Rachel M. Gillum. (2013) "Religiosity-of-interviewer effects: Assessing the impact of veiled enumerators on survey response in Egypt." *Politics and Religion* 6 (3): 459-482.

Bush, Sarah Sunn, and Lauren Prather. "Do electronic devices in face-to-face interviews change survey behavior? Evidence from a developing country." *Research & Politics* 6, no. 2 (2019): 2053168019844645.

Caeyers, Bet, Neil Chalmers, and Joachim De Weerdt. (2012). "Improving consumption measurement and other survey data through CAPI: Evidence from a randomized experiment." *Journal of Development Economics* 98, no. 1 (2012): 19-33.

Calinescu, M., Bhulai, S., & Schouten, B. (2013). Optimal resource allocation in survey designs. European Journal of Operations Research, 226(1), 115–121

Calinescu, M., Schouten, B., & Bhulai, S. (2012). Adaptive survey designs that minimize nonresponse and measurement risk. Discussion paper 201224, Statistics Netherlands

Calinescu, M., & Schouten, B. (2016). Adaptive Survey Designs for Nonresponse and Measurement Error in Multi-Purpose Surveys. Survey Research Methods, 10(1), 35–47. https://doi.org/10.18148/srm/2016.v10i1.6157

Capistrano, D. & Creighton, M. (2022). The effect of advance letters on survey participation: The case of Ireland and the European Social Survey. Survey Methods: Insights from the Field. Retrieved from https://surveyinsights.org/?p=15979.

Carletto, C., Chen, H., Kilica, T., & Perucci F. (2022). Positioning household surveys for the next decade. Statistical Journal of the IAOS. 38(1):1-24. DOI 10.3233/SJI-220042.

Centeno, Lena, Jesus Arrue, Cindy Good, Brad Edwards, and Darby Steiger. "Practical considerations for developing and implementing a video mode for survey data collection." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* (2024): 07591063241277065.

Cohen, Mollie J., and Sebastian Larrea. 2018. "Assessing and Improving Interview Quality in the 2016/17 AmericasBarometer." Methodological Insights Note #002. Nashville, TN: Vanderbilt University - LAPOP Lab. https://www.vanderbilt.edu/lapop/insights/IMN002en.pdf

Cohen L. and Lotan R. 2014. *Designing groupwork strategies for the heterogeneous classroom*, 3rd Edition. New York: Teachers College Press

Cohen, Mollie J., and Zach Warner. "How to get better survey data more efficiently." *Political Analysis* 29, no. 2 (2021): 121-138.

Conrad F. G. and Schober M. F. 2001. Clarifying survey questions when respondents don't know they need clarification. https://www.fcsm.gov/assets/files/docs/2001FCSM_Conrad.pdf

Conrad F. G. and Schober M. F. 2000. Clarifying question meaning in a household telephone survey. Public Opinion Quarterly, 64: 1-28.

Conrad F. G. and Schober M. F. 1999. A conversational approach to text-based computer administered questionnaires. In Proceedings of the Third International ASC conference. Chichester, UK: Association for Survey Computing: 91-101.

Conrad F. G. and Schober M. F. 2020. Clarifying question meaning in standardized interviews can improve data quality even though wording may change: a review of the evidence. International Journal of Social Research Methodology, 24(2), 203–226. https://doi.org/10.1080/13645579.2020.1824627

Couper, Mick P., and Edith D. De Leeuw. "Nonresponse in cross-cultural and cross-national surveys." *Cross-cultural survey methods* (2003): 157-177.

Dabalen, Andrew, Alvin Etang, Johannes Hoogeveen, Elvis Mushi, Youdi Schipper, and Johannes von Engelhardt. (2016). *Mobile phone panel surveys in developing countries: a practical guide for microdata collection*. World Bank Publications.

Daikeler J. and Bosnjak M. 2020. General Interviewing Techniques: Developing Evidence-Based Practices for Standardized Interviewing, in Olson K., Smyth J. D., Dykema J., Holbrook A. L., Kreuter F., and West B. T., *Interviewer Effects from a Total Survey Error Perspective*. Boca Raton: Taylor & Francis Group: 47-62.

Daikeler J, Bosnjak M. How to Conduct Effective Interviewer Training: A Meta-Analysis and Systematic Review. In: *Interviewer Effects from a Total Survey Error Perspective*. Vol 1. 1st ed. CRC Press; 2020:47-60. doi:10.1201/9781003020219-6

DeBell M, Amsbary M, Brader T, et al. (2022) *Methodology Report for the ANES 2020 Time Series Study*. Palo Alto, CA, and Ann Arbor, MI: Stanford University and the University of Michigan. https://electionstudies.org/wp-content/uploads/2022/08/anes_timeseries_2020_methodology_report.pdf

De Jong, J. (2016). Data collection: general consideration. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved August, 5, 2024, from http://ccsg.isr.umich.edu/.

De Leeuw, Edith D., Hox, Joop J., Dillman, Don A. 2008. International Handbook of Survey Methodology. New York: Routledge.

Dillman, D. A., Smyth, J. D and Christian, L. M. (2009) "Internet, Mail and Mixed-Mode Surveys: the Tailored Design Method, Third Edition", John Wiley and Sons Inc, Hoboken New Jersey

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed mode surveys: The tailored design method (4th ed.). Hoboken, NJ: Wiley.

Doušak, May, Roberto Briceno-Rosas, and Joost Kappelhof. "Virtual surrounding impression: ethical and privacy-respecting tracking of face-to-face computer-assisted personal interview location." *Teorija in Praksa* 61, no. 3 (2024): 669-686.

Edwards P, Cooper R, Roberts I, Frost C. Meta-analysis of randomised trials of monetary incentives and response to mailed questionnaires. J Epidemiol Community Health. 2005 Nov;59(11):987-99. doi: 10.1136/jech.2005.034397. PMID: 16234429; PMCID: PMC1732953.

European Social Survey (2024). ESS Round 12 Survey Specification for ESS ERIC Member, Observer and Guest Countries. retrieved from Survey Specification | European Social Survey, last accessed 1 of august 2024.

Fazzi G., Martire F., Pitrone M.C., (2009) Hanging by a Thread: the Telephone Interviewers Tell (Their) Story, AAPOR conference, http://www.asasrms.org/Proceedings/y2009/Files/400015.pdf

Fazzi, Gabriella, Murgia, Manuela, Nuccitelli, Alessandra, Rossetti, Francesca, Parisi, Valentino, Piergiovanni, Roberta, Arlotta, Luigi, Giacummo, Maura. 2022. "A paradata-driven statistical approach to improve fieldwork monitoring: the case of the Non-Profit Institutions census." In Di Bella, E., Fabbris, L., Lagazio, C. (Eds.) ASA 2022 Data-Driven Decision Making – Short Papers

Firchow, Pamina, and Roger Mac Ginty. "Including hard-to-access populations using mobile phone surveys and participatory indicators." *Sociological Methods & Research* 49, no. 1 (2020): 133-160.

Fowler F. J., and Mangione T. W. 1990. Standardized Survey Interviewing. Minimizing Interviewer-Related Error. Newbury Park: Sage Publications.

Giuliani, Gianluca, Grassia, Maria Gabriella, Quattrociocchi, Luciana, Ranaldi, Rita. 2004. “New Methods for Measuring Quality Indicators of ISTAT’s New CAPI/CATI Labour Force Survey.” in Davison A. and Sardy S. (eds.), Proceedings of Q2004 European Conference on Quality in Survey Statistics: pp. 1-26.

Giuliani, G., Grassia, M. G., Quattrociocchi, L., & Ranaldi, R. (2004). New Methods for Measuring Quality Indicators of ISTAT’s New CAPI/CATI Labour Force Survey. In Davison A. and Sardy S. (eds.), Proceedings of Q2004 European Conference on Quality in Survey Statistics: pp. 1-26.

Goel, Deepti, and Rosa Abraham. "Improving survey quality using paradata: Lessons from a field survey in India." *Development Policy Review* (2024): e12813.

Goel, D., Abraham, R., & Lahoti, R. (2021). Improving Survey Quality Using Para Data: Lessons from the India Working Survey. Institutional Research Grant Report Number 03, Report submitted to NCAER National Data Innovation Centre, New Delhi: National Council of Applied Economic Research.

Goodrich, Brittney, Marieke Fenton, Jerrod Penn, John Bovay, and Travis Mountain. "Battling bots: Experiences and strategies to mitigate fraudulent responses in online surveys." *Applied Economic Perspectives and Policy* 45, no. 2 (2023): 762-784.

Gourlay, Sydney, Talip Kilic, Antonio Martuscelli, Philip Wollburg, and Alberto Zezza. "High-frequency phone surveys on COVID-19: good practices, open questions." Food policy 105 (2021): 102153.

Grassi D. and Romano M.C. 2020. *L'approccio trasversale alla formazione delle reti di rilevazione* (eds.) Istat: E-Book. pp. 7-18.

Greenleaf, Abigail R., Shani R. Turke, Fiacre Bazié, Nathalie Sawadogo, Georges Guiella, and Caroline Moreau. "Interviewing the interviewers: Perceptions of interviewer–respondent familiarity on survey process and error in Burkina Faso." *Field Methods* 33, no. 2 (2021): 107-124.

Groves, Robert M., and Mick P. Couper. *Nonresponse in household interview surveys*. John Wiley & Sons, 2012.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169, 3, 439-457.

Groves R. M. and McGonagle K.A,. 2001. A theory-guided interviewer training protocol regarding survey participation. Journal of official statistics, 17(2):249–265.

Hasanbasri, Ardina, Talip Kilic, Gayatri Koolwal, and Heather Moylan. "Using paradata to assess respondent burden and interviewer effects in household surveys: Evidence from low-and middle-income countries." *Statistical Journal of the IAOS* 40, no. 2 (2024): 247-267.

Hinda, Marija, Tolić, Nebojša (Statistical Office of the Republic of Serbia [SORS]). 2021. “Data collection tool for survey monitoring and analysis.” In Conference of European Statisticians – Expert Meeting On Statistical Data Collection.

Holbrook AL, Johnson TP, Krysan M. Race-and Ethnicity-of-Interviewer Effects. Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment. 2019; p. 197–224.

Hui, C. H. and Triandis, H. C. (1989). Effects of culture and response format on extreme response style. Journal of Cross-Cultural Psychology, 20,3, 296-309.

Iarossi Giuseppe. 2006. The Power of Survey Design, The International Bank for Reconstruction and Development. The World Bank: Washington, DC

ISWGHS 2024 – Survey circulated to chapter authors; need guidance on how to cite it.

Jabkowski, Piotr. (2017). A meta-analysis of within-household selection impact on survey outcome rates, demographic representation and sample quality in the European Social Survey. *ASK. Research & Methods*, (26), 31-60.

Jäckle, Annette, Alexander Wenz, Jonathan Burton, and Mick P. Couper. "Increasing participation in a mobile app study: the effects of a sequential mixed-mode design and in-interview invitation." *Journal of Survey Statistics and Methodology* 10, no. 4 (2022): 898-922.

Jahun, I., Greby, S.M., Adesina, T., Agbakwuru, C., Dalhatu, I., Yakubu, A., Jelpe, T., Okoye, M., Ikpe, S., Ehoche, A., Abimiku, A., Aliyu, G., Charurat, M., Greenwell, G., Bronson, M., Patel, H., McCracken, S., Voetsch, A.C., Parekh, B., Swaminathan, M., Adewole, I., Aliyu, S. (2021). Lessons from rapid field implementation of an hiv population-based survey in Nigeria, 2018. JAIDS Journal of Acquired Immune Deficiency Syndromes [Internet]; 87(1): S36–42. Available from: https://journals.lww.com/10.1097/QAI.0000000000002709

Jans, M., Sirkis, R. and Morgan, D. (2013). Managing Data Quality Indicators with Paradata Based Statistical Quality Control Tools: The Keys to Survey Performance. In Improving Surveys with Paradata, F. Kreuter (Ed.). https://doi.org/10.1002/9781118596869.ch9

Jones, N. and Lewis, D. (eds, with Aitken, A.; Hörngren, J. and Zilhão M. J.) (2003). Handbook on improving quality by analysis of process variables. Final Report, Eurostat

Kam, Cindy D., and Jennifer M. Ramos. "Joining and leaving the rally: Understanding the surge and decline in presidential approval following 9/11." *Public Opinion Quarterly* 72, no. 4 (2008): 619-650.

Kappelhof, J. (2015a). *Surveying ethnic minorities: The impact of survey design on data quality.* The Hague: Sociaal en Cultureel Planbureau. 2015-5. Published on the website of the Netherlands Institute for Social Research, University of Utrecht. Retrieved from Surveying ethnic minorities (scp.nl), Accessed 5 August 2024.

Kappelhof, J.W.S. (2015b). The Impact of Face-to-Face versus Sequential Mixed Mode Designs on the Possibility of Nonresponse Bias in Surveys among non-Western minorities in The Netherlands. Journal of Official Statistics, 31, 1, 1-31.

Kappelhof, J. (2017). Survey research and the quality of survey data among ethnic minorities. Total survey error in practice, 235-252.

Kappelhof, J.W.S. and de Leeuw, E.D. (2019). Estimating the impact of measurement differences introducing by efforts to reach a balanced response among non-western minorities. *Sociological Methods & Research* **48** (1): 116–155.

Kiecker, Pamela, and James E. Nelson. "Do interviewers follow telephone survey instructions?." *Market Research Society. Journal.* 38, no. 2 (1996): 1-14.

Knowles M., Holton III E. F., Robinson, P. A. and Swanson R. A. 2005. The adult learner: The definitive classic in adult education and human resource development (6th ed.). Burlington, NJ: Elsevier.

Kołczyńska, Marta, Piotr Jabkowski, Stephanie Eckman, Interviewer Involvement in Respondent Selection Moderates the Relationship Between Response Rates and Sample Bias in Cross-National Survey Projects in Europe, *Journal of Survey Statistics and Methodology*, Volume 12, Issue 1, February 2024, Pages 1–13, https://doi.org/10.1093/jssam/smad013

Kreuter, F., Couper, M.P., & Lyberg, L.E. (2010). The use of paradata to monitor and manage survey data collection. In *Proceedings of the joint statistical meetings, American Statistical Association*, pp. 282-296.

Kreuter, Frauke, Mick Couper, and Lars Lyberg. (2010) "The use of paradata to monitor and manage survey data collection." In *Proceedings of the joint statistical meetings, American Statistical Association*, pp. 282-296.

Kreuter, Frauke, ed. 2013. *Improving surveys with paradata: Analytic uses of process information*. John Wiley & Sons.

Kreuter, Frauke, Couper, Mick, Lybergz, Lars. 2010. "The use of paradata to monitor and manage survey data collection." Joint Statistical Meetings (JSM) – Section on Survey Research Methods

Kreuter, F., and Olson, K. (2013). Paradata for Nonresponse Error Investigation. In F. Kreuter (Ed.), Improving Surveys with paradata. Analytic Uses of Process Information (pp. 13-42). Wiley Series in Survey methodology. Hoboken, New Jersey: John Wiley & Sons.

Krosnick J. A., Narayan S., Smith W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Program Evaluation*, 70, 29-44. doi: 10.1002/ev.1033

Kumar AD, Durham DD, Lane L, Perera P, Rivera MP, Henderson LM. Randomized control trial of unconditional versus conditional incentives to increase study enrollment rates in participants at increased risk of lung cancer. J Clin Epidemiol. 2022 Jan;141:11-17. doi: 10.1016/j.jclinepi.2021.08.027. Epub 2021 Aug 29. PMID: 34469801; PMCID: PMC8903037.

Kunz, Tanja, Jessica Daikeler, and Daniela Ackermann-Piek. "Interviewer-Observed Paradata in Mixed-Mode and Innovative Data Collection." *International Journal of Market Research* 66, no. 1 (2024): 14-26.

Kutka, A., Durazo, J. McGee, K.; Shaw, J. A. 2023. *"A Practical Guide to Fieldwork Training"*. LSMS Guidebook Washington, D.C. : World Bank Group. https://lsms-worldbank.github.io/pg2sq-training.

Laflamme, Francois, and James Wagner. "Responsive and adaptive designs." *The SAGE handbook of survey methodology* (2016): 397-408.

Lasky-Fink, Jessica and Rogers, Todd, Revisiting the Effect of Conditional and Unconditional Incentives on Mail Survey Response Rates (August 1, 2020). HKS Working Paper No. RWP20-024, Available at SSRN: https://ssrn.com/abstract=3671024 or http://dx.doi.org/10.2139/ssrn.3671024

Lau, Charles Q., Ansie Lombaard, Melissa Baker, Joe Eyerman, and Lisa Thalji. "How representative are SMS surveys in Africa? Experimental evidence from four countries." *International Journal of Public Opinion Research* 31, no. 2 (2019): 309-330.

Lipps, O., Herzing, J. M. E., Pekari, N., Ernst Stähli, M. Pollien, A., Riedo, G., Reveilhac, M. (2019). Incentives in surveys. FORS Guide No. 08, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2019-00008

Logan, Carolyn, Pablo Parás, Michael Robbins, and Elizabeth J. Zechmeister. "Improving data quality in face-to-face survey research." *PS: Political Science & Politics* 53, no. 1 (2020): 46-50.

Lundquist, P., Sarndal, C.E. (2013). Aspects of responsive design with applications to the Swedish Living condition survey. Journal of Official Statistics; 29(4): 557–58

Lyberg, Lars E., and Paul P. Biemer. "Quality assurance and quality control in surveys." In *International handbook of survey methodology*, pp. 421-441. Routledge, 2012.

Lynn, P. (2017) "From standardised to targeted survey procedures for tackling non-response and attrition". In *Survey Research Methods*, 11(1): 93-103. http://dx.doi.org/10.18148/srm/2017.v11i1.6734.

Lynn, Peter. (2020). "Evaluating push-to-web methodology for mixed-mode surveys using address-based samples." In *Survey Research Methods* 14(1):19-30.

Lynn, P., Thomson, K. and Brook, L. (1998) "An experiment with incentives on the British Social Attitudes Survey," In *Survey Methods Newsletter* 18(2): 12-14.

Lynn, P., Nandi, A., Parutis, V. & Platt, L. (2018) 'Design and implementation of a high quality probability sample of immigrants and ethnic minorities: Lessons learnt'. *Demographic Research*, 38: 513-548. https://doi.org/10.4054/DemRes.2018.38.21.

Lynn, P., Bianchi, A. and Gaia, A. (2024) The impact of day of mailing on web survey response rate and response speed, In *Social Science Computer Review*, 42(1): 352-368. https://doi.org/10.1177/08944393231173887

Martin, Jean. "PAPI to CAPI: the OPCS experience." In *Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London*, pp. 96-117. 1993.

Menold, Natalja, Uta Landrock, Peter Winker, Nathalie Pellner, and Christoph J. Kemper. "The impact of payment and respondents' participation on interviewers' accuracy in face-to-face surveys: Investigations from a field experiment." *Field Methods* 30, no. 4 (2018): 295-311.

Mercader, Hannah Faye G., Jerome Kabakyenga, David Tumusiime Katuruba, Amy J. Hobbs, and Jennifer L. Brenner. (2017) "Female respondent acceptance of computer-assisted personal interviewing (CAPI) for maternal, newborn and child health coverage surveys in rural Uganda." *International journal of medical informatics* 98: 41-46.

Metre, K. V., Mathur, A., Dahake, R., Bhapkar, Y., Ghadge, J., Jain, P., & Gore, S. (2024). An Introduction to Power BI for Data Analysis. International Journal of Intelligent Systems and Applications in Engineering. 142-147.

Montalvo, J. Daniel, Mitchell A. Seligson, and Elizabeth J. Zechmeister. 2018. "Improving Adherence to Area Probability Sample Designs: Using LAPOP's Remote Interview Geo-locating of Households in real-

Time (RIGHT) System." Insights Methodological Note #004. Nashville, TN: Vanderbilt University - LAPOP Lab. https://www.vanderbilt.edu/lapop/insights/IMN004en.pdf

Morren, M., Gelissen, J. P., and Vermunt, J. K. (2012). Response Strategies and Response Styles in Cross-Cultural Surveys. Cross-Cultural Research, 46, 3, 255-279.

Morton-Williams, Jean. (1993), Interviewer Approaches, Dartmouth Publishing, Aldershot.

Mneimneh, Zeina N., Julie De Jong, Kristen Cibelli Hibben, and Mansoor Moaddel. (2020). "Do I look and sound religious? Interviewer religious appearance and attitude effects on respondents' answers." *Journal of Survey Statistics and Methodology* 8 (2): 285-303.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

NSB – National Statistics Bureau. 2018. Guidelines for conduct of surveys. Royal Government of Bhutan, National Statistics Bureau, Thimphu.

O'Hare, B. C., Jans, M. E. (2011). Designing a paradata application in a CAPI environment.

Olson, Kristen, and Ipek Bilgen. "The role of interviewer experience on acquiescence." *Public opinion quarterly* 75, no. 1 (2011): 99-114. https://doi.org/10.1093/poq/nfm007.

Olson, Kristen, and Andy Peytchev. "Effect of interviewer experience on interview pace and interviewer attitudes." *Public opinion quarterly* 71, no. 2 (2007): 273-286. https://doi.org/10.1093/poq/nfm007.

Oseni, Gbemisola, Palacios-Lopez, Amparo, Mugera, Harriet Kasidi and Durazo, Josefine. 2021. Capturing What Matters: Essential Guidelines for Designing Household Surveys. Washington DC: World Bank.

Pforr, K. (2016). Incentives. GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. doi: 10.15465/gesis-sg_en_001.

Pinzón, Natalia, Vikram Koundinya, Ryan Galt, William Dowling, Marcela Boukloh, Namah C. Taku-Forchu, Tracy Schohr et al. "AI-powered Fraud and the Erosion of Online Survey Integrity: An Analysis of 31 Fraud Detection strategies." Frontiers Vol. 9 doi: 10.3389/frma.2024.1432774

Randall, Sara, Ernestina Coast, Natacha Compaore, and Philippe Antoine. 2013. "The power of the interviewer." *Demographic research* 28: 763-792.

Rossetti, F., Bontempi, K. & Pietropaoli, S. A comparison between Video Mediated Interview (VMI) and CAPI in Labour Force Survey. METRON (2024). https://doi.org/10.1007/s40300-024-00282-7

Roll-Hansen D. and Rustenbach E. 2021. Sustainable statistical training programs at National Statistical Offices. New York: Global Network of Institutions for Statistical Training (GIST) https://unstats.un.org/gist/resources/documents/Sustainable-statistical-training-programs-at-NSOs.pdf

Sand G., Jeny, T.P., Yuri P. 2019. Interviewer Training, in Bergmann M., Scherpenzeel, A., Börsch-Supan A. (Eds.) Share Wave 7 Methodology: Panel Innovations and Life Histories, Munich Center for the Economics of Aging (MEA) at the Max Planck Institute for Social Law and Social Policy (MPISOC): 89-97

Schenk, Patrick Oliver, and Simone Reuß. 2024. "Paradata in surveys." In *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge*, pp. 15-43. Cham: Springer International Publishing.

Schnell R. and Trappmann M. 2006. The effect of the refusal avoidance training experiment on final disposition codes in the German ESS-2. Working Paper, Center for Quantitative Methods and Survey Research, University of Konstanz: https://www.unidue.de/~hq0215/documents/2006/2006EffectRefusalAvoidanceTraining.pdf

Schouten, B., Calinescu, M., Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. Survey Methodology, 39 (1), 29 – 58.

Schouten, Barry, Fannie Cobben, and Jelke Bethlehem. "Indicators for the representativeness of survey response." *Survey methodology* 35, no. 1 (2009): 101-113.

Schouten, B., Peytchev, A., & Wagner, J. (2017). Adaptive Survey Design (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315153964

Shah, Neha, Diwakar Mohan, Jean Juste Harisson Bashingwa, Osama Ummer, Arpita Chakraborty, and Amnesty E. LeFevre. "Using machine learning to optimize the quality of survey data: protocol for a use case in India." *JMIR Research Protocols* 9, no. 8 (2020): e17619.

Signore, Marina, Brancato, Giovanna, Carbini, Riccardo, D'Orazio, Marcello, Simeoni, Giorgia. 2012. Quality Guidelines for Statistical Processes. Roma: Italian National Institute of Statistics (Istat).

Singer, Eleanor, John Van Hoewyk, Nancy Gebler, and Katherine McGonagle. "The effect of incentives on response rates in interviewer-mediated surveys." *Journal of official statistics* 15, no. 2 (1999): 217.

Singer, Eleanor and Cong Ye. "The Use and Effects of Incentives in Surveys." Annals of the American Academy of Political and Social Science 645, no. 1 (2013): 112-141.

SRC – Survey research center. 2011. Guidelines for Best Practice in Cross-Cultural Surveys. FULL GUIDELINES. Institute for Social Research, University of Michigan.

Statistics Canada. 2003. Survey Methods and Practices. Catalogue no. 12-587-X2003001

Stiegler A. and Biedinger N. 2016. Interviewer Skills and Training. GESIS Survey Guidelines. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.15465/sdm-sg013.

Stoop, I. A. L. (2005). The hunt for the last respondent. Nonresponse in sample surveys. The Hague: Sociaal en Cultureel Planbureau

Stoop, I. A., Billiet, J., Koch, A., & Fitzgerald, R. (2010). Improving survey response: Lessons learned from the European Social Survey. John Wiley & Sons.

Struwig, Jarè, and Benjamin J. Roberts. "Truth be in the field: conducting comparative survey research in a multicultural society such as South Africa." *International Journal of Sociology* 50, no. 2 (2020): 142-162.

Tagaram, S.D., Chen, C. (2024). Quality Tools and Techniques (Fishbone Diagram, Pareto Chart, Process Map). In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan–. PMID: 39383290. Available from: https://www.ncbi.nlm.nih.gov/books/NBK607994/

Thornberry, Owen, Benjamin Rowe, and Ronald Biggar. "Use of CAPI with the US National Health Interview Survey." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 30, no. 1 (1991): 27-43.

Tourangeau, R. ed., 2014. *Hard-to-survey populations*. Cambridge University Press.

Tourangeau, Roger, J. Michael Brick, Sharon Lohr, and Jane Li. "Adaptive and responsive survey designs: a review and assessment." *Journal of the Royal Statistical Society Series A: Statistics in Society* 180, no. 1 (2017): 203-223. https://doi.org/10.1111/rssa.12186.

Tourangeau R., Rasinski K. A. (1988). Cognitive processes underlying context effects in attitude measurement. Psychological Bulletin, 3, 299-314. doi: 10.1037//0033-2909.103.3.299

Wagner, J. (2008). Adaptive Survey Design to Reduce Nonresponse Bias. PhD thesis, University of Michigan.

Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. Public Opinion Quarterly 74, 2, 223-243.

Walsh, R., Coombs, J. (2013). Case Reassignment: When making contact is a two-person job. In Joint Statistical Meetings (JSM) 2014 - Survey Research Methods Section

Weinreb, Alexander, Mariano Sana, and Guy Stecklov. "Strangers in the field: A methodological experiment on interviewer–respondent familiarity." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 137, no. 1 (2018): 94-119.

West, Brady T. 2011. "Paradata in Survey Research." *Survey Practice* 4 (4). https://doi.org/10.29115/SP-2011-0018.

Winker, Peter, Karl-Wilhelm Kruse, Natalja Menold, and Uta Landrock. "Interviewer effects in real and falsified interviews: Results from a large scale experiment." *Statistical Journal of the IAOS* 31, no. 3 (2015): 423-434.

Wuyts, C., Loosveldt, G. (2020). Measurement of Interviewer Workload within the Survey and an Exploration of Workload Effects on Interviewers' Field Efforts and Performance. Journal of Official Statistics, vol. 36, no. 3, Sciendo, pp. 561-588. https://doi.org/10.2478/jos-2020-0029

Xiao, Ziang, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. "Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions." *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, no. 3 (2020): 1-37.

Zimbalist, Zack. (2022) "Bystanders and response bias in face-to-face surveys in Africa." *International Journal of Social Research Methodology* 25 (3): 361-377.

This page has been intentionally left blank.

CHAPTER 7
DATA PROCESSING

This is a placeholder page.

Chapter 7 will be completed in Spring 2025.

This page has been intentionally left blank.

CHAPTER 8
WEIGHTING

Diego Zardetto (The World Bank)
Kevin McGee (The World Bank)

CHAPTER 8
WEIGHTING

8.1. Introduction

1. Weighting is a crucial phase in the production process of probability sample surveys. During this phase, numeric expansion factors, called survey weights, are calculated for all the surveyed analysis units. These weights are then used to produce estimates of the target population. The analysis task (see Chapter 9) is accomplished through weighted estimators, where the values of the interest variables observed on each analysis unit are multiplied by the survey weight of the unit. Most surveys of households and individuals implement a complex sampling design (see Chapter 5) and, as such, weights are essential for accounting for differences in the inclusion probabilities of the sampled units. Because survey weights are instrumental in making inferences about the target population based on the observed sample, they significantly influence the accuracy of survey estimates. Therefore, survey weights must be regarded as an integral component of the survey data and a critical factor affecting its quality.

2. This chapter addresses the weight calculation phase of sample surveys. It is organized as follows:

- **Section 8.1** introduces the concept and purpose of survey weights, outlining the weight calculation process in real-world surveys with a focus on fundamental steps, objectives, and statistical principles. While emphasizing statistical reasoning and terminology, the discussion avoids mathematical formalism and technical implementation details.
- **Section 8.2** provides a practical guide to implementing a weight calculation process, following the typical sequence of steps used in real-world applications and including basic mathematical descriptions of key statistical techniques employed at each step.
- **Section 8.3** covers specialized topics, presenting weighting considerations for panel surveys and pooled survey samples.
- **Section 8.4** explores emerging approaches, illustrating how established weighting methodologies can be adapted to achieve unbiased estimation from nonprobability samples.

8.1.1. Objectives in calculating survey weights

3. When calculating survey weights, the key objective is to construct estimators with desirable statistical properties. Above all, good estimators should be unbiased, or nearly unbiased. Substantial bias leads, on average, to poor estimates and prevents the construction of valid confidence intervals. In addition, good estimators should be as efficient as possible (i.e. minimize sampling variance). A nearly unbiased estimator with small sampling variance is likely to produce estimates that are close to the true value of the population parameter. These criteria translate into a clear-cut hierarchy of requirements for the construction of survey weights:

- **Achieve unbiasedness under ideal conditions**: Under ideal conditions (i.e. in the absence of nonsampling errors, such as undercoverage, nonresponse, and measurement errors), survey weights should ensure the unbiasedness of survey estimates.
- **Minimize bias in real-world surveys**: In real-world surveys, survey weights should minimize biases induced by nonsampling errors (e.g. nonresponse bias and frame undercoverage bias).

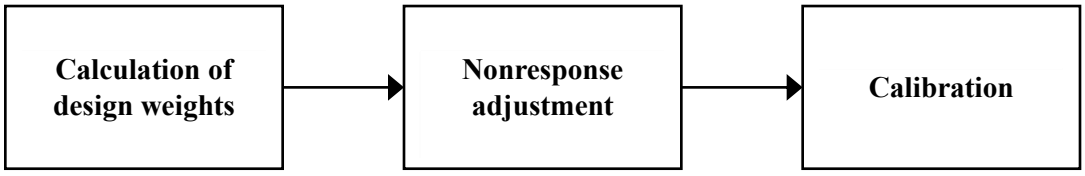
- **Leverage auxiliary information to increase accuracy**: When auxiliary information on the target population is available from external sources, survey weights should leverage that information to reduce bias and/or improve the precision of survey estimates.

8.1.2. Weight calculation process

4. In all but the rarest cases, it is not possible to simultaneously meet all three requirements in a single calculation step. Instead, survey weights are typically developed through a sequential multistep process, where each step refines the weights from the previous step before passing the adjusted weights to the next calculation step. The overarching principle is that the quality of the weights should progressively improve throughout the weighting process. In other words, each step in the sequence should enhance the weights it receives without compromising the gains accumulated in previous steps. To this end, weight adjustments need to be principled, controlled, and minimal. This means that each weighting adjustment step must (i) aim at a clearly stated statistical goal (principled), (ii) use established and mathematically rigorous statistical methods (controlled), and (iii) achieve its goal while modifying the input weights as little as possible (minimal). Conversely, heuristic or subjective modifications of the weights should be avoided, as they are likely to damage the inferential quality of the survey.

5. In practice, weight calculation processes can be quite elaborate, involving many steps and varying from survey to survey. Differences across surveys may depend on the statistical domain of interest (e.g. household surveys vs. enterprise surveys), the type of survey (e.g. cross-sectional vs. longitudinal), the sampling design (e.g. one-stage unit sampling designs vs. multi-stage cluster sampling designs), the availability of auxiliary information (e.g. from Censuses, sampling frames, administrative data, or other surveys), and – crucially – the nonsampling errors that the weighting process is designed to address and mitigate (e.g. nonresponse or frame imperfections). Despite these variations, the weighting processes of most surveys share three common major steps shown in **Figure 8.1**.

Figure 8.1. Three core steps in the weighting process.



8.1.3. Calculation of design weights

6. The design weights calculated in the first step form the cornerstone of the entire weighting process. For this reason, they are also often referred to as “base weights” or “initial weights”. As their name suggests, the design weights are directly and solely informed by the sampling design of the survey (see Chapter 5). Specifically, the design weight of a unit is the reciprocal of the unit’s sample inclusion probability. Historically, this definition originates from the mathematical expression of the Horvitz-Thompson estimator of population totals. Since the Horvitz-Thompson estimator is known to be unbiased under the sampling design, the design weights are sufficient to achieve estimation unbiasedness under ideal conditions, thereby meeting the first fundamental requirement of the weighting process.

7. Intuitively, the mechanism through which the design weights secure this essential property is straightforward. Design weights expand the values observed in the selected sample in such a way that units that were less likely to be sampled have their values magnified more than units that were more likely to be

sampled. This differential expansion perfectly compensates in estimation for any differences in inclusion probabilities across the population units. Since most surveys use complex sampling designs, which result in unequal inclusion probabilities, the design weights must be considered the necessary starting point of any further weighting adjustment. Unweighted estimators (e.g. the sample mean) should be avoided when analyzing complex survey data because they are generally biased, even in the absence of nonsampling errors.

8. Thanks to the principles of probability sampling, inclusion probabilities are strictly positive and known. In particular, the design weights are calculable, finite, and greater or equal to one. These properties underpin the popular interpretation of the design weight as "the number of population units the sample unit represents"[1]. If, for instance, a simple random sample of 20 units is selected without replacement from a population of 100 units, then the inclusion probability of each population unit is 0.2 (0.2 = 20 / 100). It follows that the design weight of the 20 selected units is 5 (5 = 1 / 0.2), which can be intuitively interpreted to mean that each unit in the sample "represents" 5 units in the target population. Consistent with this interpretation, the sum of the design weights over the sample exactly yields the population size, namely 100 (100 = 5 * 20). Note that, for a generic sampling design, the sum of the design weights over the sample is an unbiased estimator of the population size but does not need to exactly match the true (and typically unknown) value of the population size.

8.1.4. Beyond design weights: Weighting adjustments

9. Once the design weights have been produced, they can be passed on to the subsequent steps of the weighting process. There are essentially two reasons to adjust the design weights. Under ideal conditions, namely in the absence of nonsampling errors, the only meaningful reason would be to increase the precision of the survey estimates, since the design weights already ensure unbiasedness. The desire for increased estimate precision motivates the search for estimators that are more efficient than the Horvitz-Thompson estimator but still nearly unbiased. Historically, this line of research culminated in the development of calibration estimators. Calibration estimators improve precision by using auxiliary information on the target population that is available from external sources. Nowadays, most surveys apply calibration adjustments to the weights to varying degrees. The form of these adjustments depends on the used auxiliary information, a topic that will be discussed in Section 8.1.6.

10. In real-world surveys, the main reason to adjust the design weights is to mitigate biases caused by nonsampling errors (e.g., due to nonresponse and to undercoverage). While the randomness arising from probability sampling is fully controlled by the statistician, nonsampling errors introduce uncontrolled randomness and systematic effects that are unknown. Because of these unknown selective forces, bias can arise, and the design weights are no longer sufficient to guarantee unbiased survey estimates.

11. A general approach to overcome this issue consists of describing the nonsampling errors affecting the data through a statistical model, and then using the predictions from the fitted model to compensate for the biasing effects of the nonsampling errors. In practice, it is the fitted model that determines the form of the weight adjustment. Therefore, the success of the weight adjustment in reducing bias depends on the validity of the model. Furthermore, each source of nonsampling error may require a different statistical

[1] While undoubtedly intuitive and useful, this interpretation should be limited to design weights and not automatically extended to general survey weights. Indeed, there exist a whole class of estimators, known as calibration estimators, which employs weights that are more complex than design weights and can (legitimately) be less than 1 or even negative.

model and hence a dedicated step in the weighting process. The next two sections introduce two core steps to address nonsampling errors in real-world surveys: non-response adjustment and calibration.

8.1.5. Nonresponse adjustment

12. Nonresponse is arguably the most widespread source of nonsampling error since it affects virtually every real-world survey. For this reason, weighting processes almost always include a nonresponse adjustment step. Survey statisticians distinguish between two kinds of nonresponse error: unit (or total) nonresponse and item nonresponse. Unit nonresponse occurs when a sampled unit, for whatever reason, either does not respond at all to the survey, or fails to provide enough information for its data to be usable at all in the estimation phase. As a result, the nonrespondent unit cannot contribute to any survey estimate. Item nonresponse happens when a respondent unit partially completes the questionnaire, thus generating missing data only for specific items. In this case, the unit can still contribute to some survey estimates. Weighting adjustments can only address unit nonresponse, whereas the treatment of item nonresponse is usually addressed via editing and imputation techniques[2]. Therefore, in the rest of this chapter, the term "nonresponse" is meant to be understood as "unit nonresponse".

13. Unless countermeasures are taken, nonresponse has two detrimental effects: it makes survey estimates biased and less precise. The main cause of efficiency loss is the reduction in sample size caused by nonresponse. Additionally, differential nonresponse rates across sampling strata can jeopardize the planned allocation of the sample, making the sampling design even less efficient. However, the loss in precision is usually considered a minor disturbance compared to the threat posed by nonresponse bias. Accordingly, nonresponse weighting adjustments are primarily aimed at gaining protection against nonresponse bias.

14. Bias arises from nonresponse when subpopulations characterized by different response rates also happen to differ with respect to the variables of interest for the survey. For instance, in a hypothetical survey on smoking habits, if males respond less frequently than females and males tend to smoke more cigarettes per day than females, then the survey estimate of average cigarettes smoked per day would be underestimated, unless the design weights are adjusted to compensate for nonresponse. In other words, the use of plain design weights would lead to downward biased estimates because males, who tend to smoke more than females, would be underrepresented in the respondent sample due to their lower propensity to respond. Intuitively, to offset the bias, one could apply a nonresponse adjustment such that the design weights of male respondents are inflated more than those of female respondents. To this end, it would be reasonable to multiply the design weights of males and females by the reciprocal of their respective response rates, thereby compensating for the underrepresentation of males among the respondents. While these ideas provide a good intuition of how nonresponse adjustments are implemented in practice, the illustrated example oversimplifies the main challenge with nonresponse: identifying variables that explain the response behavior of the units well (in the example, sex).

15. In real-world surveys, statisticians often formulate and fit an explicit model that tries to explain the units' response behavior (i.e. whether they responded or not) in terms of all the variables that are known for both respondents and nonrespondents[3]. In this respect, logistic regression is a popular modeling choice. The

[2] This is unsurprising, since unit nonresponse is unit specific as survey weights are, whereas item nonresponse is variable specific.

[3] Since little is typically known about nonrespondents, it is often the case that a rather limited set of such variables exist. In lucky cases where variables are abundant, the fitted model can be analysed to retain just a parsimonious set of powerful variables.

fitted model is then used to estimate the response propensity (i.e. probability) of each respondent unit based on its profile of explanatory variables. Lastly, the reciprocals of the estimated response propensities are used to expand the design weights. Since estimated response propensities are positive and less or equal to 1, all the design weights either stay the same or receive an upward adjustment. This is in line with intuition, as nonresponse causes the realized sample to be smaller than planned, so that the original design weights would now be too small for the respondent units to represent the whole population. Additionally, respondent units with lower estimated response propensity receive larger upward adjustments. Overall, the structure of the nonresponse adjustment factors mimics that of the design weights, with the crucial difference that inclusion probabilities were controlled and known, whereas response propensities are uncontrolled and unknown and thus need to be estimated from the sample. This also explains why bias mitigation critically depends on the explanatory power of the propensity model variables.

16. Once the best possible nonresponse adjustments have been computed and applied to the design weights of the respondent sample, the resulting adjusted weights provide protection against nonresponse bias and can be passed on to the subsequent weighting step.

8.1.6. Calibration

17. Most large-scale surveys conducted by National Statistical Offices (NSO) implement at least one calibration step to further refine the nonresponse-adjusted weights. Calibration is a methodology to improve the quality of survey estimates by using, in a systematic and rigorous way, auxiliary information on the target population that is available from external sources. Historically, calibration was developed to enhance the precision of survey estimates under ideal conditions, but it can also reduce biases, such as undercoverage bias or nonresponse bias, depending on the nature and richness of the auxiliary information.

18. The calibration methodology generates a class of estimators known as calibration estimators, which use special weights called calibration weights. Calibration weights are obtained by minimally adjusting the input weights[4] so that survey estimates exactly match population totals known from external sources. The survey variables for which population totals are available are called auxiliary variables. The minimality of the calibration adjustment plays a critically important role. When the input weights to the calibration step have already been adjusted for nonresponse, the minimality of the adjustment avoids undermining the nonresponse bias reduction achieved by the input weights. At the same time, because the calibration weights yield perfect estimates of the population totals of the auxiliary variables, calibration results in improved estimates of any interest variables correlated with the auxiliary variables. As an additional benefit, calibration allows for the dissemination of survey estimates that are consistent with the external source. When that source is public, the achieved coherence promotes credibility in disseminated statistics.

19. Calibration typically increases estimation efficiency: the stronger the correlation between the interest variables and the auxiliary variables, the larger the efficiency gain (i.e. sampling variance reduction). In addition, calibration can mitigate undercoverage bias. When units of the target population are missing from the sampling frame, they receive an inclusion probability of zero, which marks a radical departure from the principles of probability sampling. As a result, the sample cannot represent the target population, and the design weights cannot ensure unbiasedness, even in the absence of nonresponse. Specifically, undercoverage bias arises when the units not covered by the sampling frame differ from the covered units with respect to the interest variables. For instance, in a hypothetical telephone survey on healthcare

[4] Because of its versatility, calibration is increasingly being used to deal with nonresponse, as an alternative approach to response propensity modeling. In this case, the "input weights" to the calibration step would be design weights. Otherwise, the "input weights" to the calibration step already incorporate a nonresponse adjustment.

expenditures, if poor households are more frequently missing from the list frame of phone numbers than non-poor households, and poor households tend to have lower healthcare expenditures, then the unadjusted survey estimate of average healthcare expenditure would be overestimated.

20. Calibration can reduce undercoverage bias provided that (i) the known totals originate from external sources of better quality than the sampling frame (i.e. they must have higher coverage and ideally be more current than the frame) and (ii) the auxiliary variables can explain to some extent the differences between covered and non-covered units. In the telephone survey example, recent census data would be a promising source, and calibration to auxiliary variables known to correlate with poverty (and hence healthcare expenditures) could substantially mitigate undercoverage bias. Natural candidates would be, e.g., geographic variables (such as region and rural-urban status), household size, education level and labor status of the household head, and housing variables. By forcing survey estimates to match the known totals of such variables, calibration would entirely remove any bias from those estimates, thereby at least reducing the bias affecting estimates of variables correlated with the auxiliaries (in the example, healthcare expenditure).

21. In general, no explicit closed-form expression exists to calculate calibration weights. Therefore, calibration adjustments are typically computed via iterative numerical optimization routines. Such adjustments are, of course, a function of the auxiliary information used in the calibration. In particular, calibration can both increase and decrease the nonresponse-adjusted weights, depending on the auxiliary variables' profile of the respondent units. This contrasts with nonresponse adjustments, which were necessarily greater than one, thereby always inflating the design weights. Calibration and explicit modeling of response propensities also differ in the granularity of the information they leverage to adjust the weights of the respondent units. Indeed, response propensity modeling requires variables that must be known for both respondents and nonrespondents. On the other hand, calibration can use auxiliary variables that are only known for respondents, provided their population totals are known from external sources.

22. Once the nonresponse-adjusted weights have been calibrated using auxiliary variables that either describe well how the sampling frame undercovers the target population or correlate with the main interest variables of the survey (or ideally do both), the weighting process is essentially complete. The final weights not only fully incorporate sampling design information but also minimize both nonresponse and undercoverage biases conditional on the available auxiliary information. Therefore, these final weights are ready for the calculation of survey estimates.

8.1.7. Cross-cutting considerations

23. An important principle in survey methodology is that survey weights must be "universal", meaning that the same set of weights can be used to compute estimates of any population parameters, regardless of the variables on which these parameters depend. Surveys conducted by NSOs are typically multipurpose and must produce estimates for a wide range of population parameters across different dissemination domains. Thus, developing universal survey weights (i.e. a unique set of unit-specific weights) is essential to ensure the scalability of production processes, thereby enhancing the timeliness and punctuality of disseminated statistics. Conversely, the development of separate weighting systems to estimate different parameters of the same target population is discouraged. Such an approach would be cumbersome and could jeopardize the internal consistency of the survey estimates. In sum, "universality" means that there should be a single set of weights produced for an entire survey dataset.

24. In some cases, survey data can be used to make inferences about different target populations. For instance, household surveys typically collect data on both private households and individuals residing in those households. Sometimes, the interviewed households are also used as a vehicle to collect data on units

belonging to other linked populations, such as non-farm enterprises (NFE) or agricultural plots owned by the households. In these cases, multiple sets of survey weights need to be developed, one for each analysis unit (e.g. households, individuals, NFEs, or plots). Nonetheless, the universality principle still applies, so that a single set of weights must suffice to estimate all parameters of each target population.

25. In certain situations, different analysis units can be used to estimate population parameters that, despite being unknown, must nonetheless satisfy strong logical constraints. In such cases, it is desirable to develop "integrated weights" across related analysis units, ensuring those logical constraints are also satisfied in estimation. For instance, in a household survey, one could use household-level weights to estimate (a) the number of households of size 4, and individual-level weights to estimate (b) the number of individuals living in households of size 4. The unknown true value of parameter (b) is, of course, logically constrained to be equal to the unknown true value of parameter (a) times 4. However, survey estimates will not automatically satisfy that same logical constraint unless (i) individual weights are constant within households and (ii) each individual has the same weight as the household to which they belong. Weights of this kind are called "integrated household-individual weights". Even when different nonresponse adjustments are applied to the household-level and individual-level weights, methods exist to produce integrated household-individual weights through the final calibration step of the weighting process (see Section 8.2.5.4).

26. While nonresponse and calibration adjustments are necessary to protect the survey estimates from bias, they tend to increase the variability of the survey weights. This effect can negatively affect the precision of survey estimates. Moreover, in some circumstances, the weight calculation process may generate "extreme" final weights for some analysis units, that is weights that are either negative or very large.

27. On the one hand, exceedingly large survey weights can lead to unstable estimates, especially for interest variables that are highly skewed at the population level, such as consumption expenditures or income. On the other hand, negative weights (or even weights whose value is less than 1), may challenge the interpretation of many users, as users often tend to interpret the final survey weights as they would interpret the design weights, namely as "the number of population units the sample unit represents".

28. Trimming techniques exist to post-process the survey weights and eliminate extreme weights. However, these techniques are heuristic and unfortunately do not rely on mature theory. In general, trimming survey weights introduces some amount of bias in survey estimates. This is a serious risk which should always be scrupulously balanced against the expected gains in terms of precision and interpretability. In general, it is advisable to apply trimming procedures sparingly and carefully (see Section 8.2.6.1).

8.2. Construction of survey weights

29. Section 8.1 introduced the concept and purpose of survey weights and provided an overview of the weight calculation process of real-world surveys in terms of fundamental steps, specific goals, and underlying statistical principles. This section elaborates on the three fundamental steps mentioned in Section 8.1 – calculation of design weights, nonresponse adjustment, and calibration – while expanding the scope to include additional weight adjustment steps that may be necessary in some circumstances (e.g., adjustment for unknown eligibility or trimming of survey weights). The aim is to provide a concise yet reasonably complete illustration of how to implement a weight calculation process in practice, by offering a basic description of its sequential steps.

30. An in-depth discussion of methodological choices and technical implementation details is beyond the scope of this section. For these topics, the interested reader will be referred to external resources.

General reviews about the construction of weights in probability surveys include Kalton and Flores-Cervantes (2003), Biemer and Christ (2012), Haziza and Beaumont (2017), Valliant and Dever (2018).

8.2.1. Design weights

31. The design weights (also known as "base weights" or "initial weights") are the foundation of the entire survey weighting process. Their purpose is to ensure the unbiasedness of survey estimates of population totals under ideal conditions (i.e. in the absence of nonsampling errors). The design weights are directly and solely informed by the sampling design of the survey, and their mathematical expression originates from the structure of the Horvitz-Thompson estimator of population totals (Horvitz and Thompson, 1952).

32. Consider a finite population U of size N. Each unit of the population is identified by an integer label k, such that $k = 1, \ldots, N$. Let s be a random sample selected from U via probability sampling under a specified sampling design. Denote a generic interest variable with y and the value of that variable for unit k with y_k. The unknown population total of the interest variable y, denoted by Y, is:

$$Y = \sum_U y_k \tag{8-1}$$

Note that, in this chapter, summation indexes are omitted whenever no ambiguities arise from this notational simplification (e.g. in Equation (8-1), the symbol U is implicitly assumed to mean $k \in U$).

33. The Horvitz-Thompson (HT) estimator of the population total Y is defined as follows:

$$\hat{Y}_{\text{HT}} = \sum_s \frac{y_k}{\pi_k} \tag{8-2}$$

where π_k denotes the inclusion probability of unit k, namely the probability that the sampling design generates a random sample s which includes population unit k:

$$\pi_k = \Pr(k \in s) \tag{8-3}$$

34. The design weight of unit k, denoted by d_k, is defined as the reciprocal of its inclusion probability under the sampling design:

$$d_k = \frac{1}{\pi_k} \tag{8-4}$$

By using the design weights, the HT estimator can be expressed as a weighted estimator:

$$\hat{Y}_{\text{HT}} = \sum_s d_k y_k \tag{8-5}$$

35. The principles of probability sampling guarantee that inclusion probabilities are known and strictly positive, $1 \geq \pi_k > 0$ for all $k \in U$. Thus, design weights are calculable, finite, and greater than or equal to one, $d_k \geq 1$. These properties underpin the popular interpretation of the design weight as "the number of population units the sample unit represents". For instance, if sample unit k has $d_k = 100$, it can be thought as representing 100 population units, including itself.

36. The HT estimator is unbiased under the sampling design. This crucial property is independent of the interest variable y. Therefore, the design weights in Equations (8-4) and (8-5) can be seen as "universal" expansion factors. More explicitly, the same set of weights d_k can be attached to the sample units $k \in s$ to

estimate without bias the totals of any variables. The design weights achieve this goal by compensating for any differences in inclusion probabilities that the sampling design may generate (e.g. in stratified sampling designs with disproportionate allocation).

37. The simplest interest variable is the variable whose value is equal to one for all population units. The population total of this variable is the population size, N. Using the notation 1_k to indicate the values of this variable (i.e. $1_k = 1$ for all $k \in U$), it follows from Equation (8-5) that the HT estimator of the population size is:

$$\widehat{N}_{\mathrm{HT}} = \sum_s d_k 1_k = \sum_s d_k \tag{8-6}$$

In other words, the sum over the sample of the design weights is an unbiased estimator of the population size. This key property holds for the design weights under ideal conditions and should be retained as much as possible by the final survey weights in real-world surveys.

38. In many sampling designs (e.g. the multistage stratified cluster sampling designs typically adopted in household surveys), the true population size N is unknown. However, it is often the case that a reasonably good estimate of N is available from external sources. Such an estimate can thus be exploited as a benchmark to perform a first-level screening for possible mistakes in the calculation of survey weights. In large-scale surveys, a marked discrepancy between the benchmark and the sum of the final weights over the respondent sample is indicative of flaws in the weighting process.

8.2.1.1. Simple random sampling and equal probability designs

39. Explicit formulas exist for the inclusion probabilities of the most common sampling designs, making the calculation of design weights straightforward. For instance, for simple random sampling without replacement (SRSWOR) of size n, the inclusion probabilities are $\pi_k = n/N$ for all $k \in U$. The SRSWOR is the simplest sampling design that results in equal inclusion probabilities for all population units. Sampling designs with this property are sometimes called EPSEM (Equal Probability of Selection Method). The design weights are also constant under the SRSWOR design:

$$d_k = \frac{N}{n} \tag{8-7}$$

Sampling designs that result in constant weights across all sample units are called “self-weighting”. An inferential feature of self-weighting designs is that they produce perfect estimates of the population size, N. For instance, under SRSWOR, one has that $\widehat{N}_{\mathrm{HT}} = \sum_s d_k = \sum_s (N/n) = N$, regardless of the selected sample s.

40. Systematic sampling with equal probabilities is another EPSEM design. Thanks to its simplicity and flexibility, it is widely used as an alternative to SRSWOR. For a sample of size n, it produces inclusion probabilities $\pi_k = n/N$ and design weights $d_k = N/n$. Although these expressions are identical to those for a SRSWOR design of the same size, the two designs are very different. Just to mention one key difference, systematic sampling from a deliberately ordered frame tends to produce more precise estimates $\widehat{Y}_{\mathrm{HT}}$ than SRSWOR, provided the sorting criteria result in an upward or downward trend in the interest variable y.

8.2.1.2. Stratified sampling

41. In general, stratified designs where units are sampled with equal probabilities within each stratum produce unequal inclusion probabilities across strata. Let the index h identify the L strata, such that

$h = 1, \ldots, L$. Denote the population size of stratum h by N_h, so that $\sum_{h=1}^{L} N_h = N$. Similarly, denote the sample allocation to stratum h by n_h, so that $\sum_{h=1}^{L} n_h = n$. Indicate with s_h the sample selected in stratum h, where $\bigcup_{h=1}^{L} s_h = s$. If the selection within each stratum is EPSEM, then the inclusion probability of unit k in stratum h is n_h/N_h. Therefore, the design weights for this design are:

$$d_{hk} = \frac{N_h}{n_h} \quad \text{for all } k \in s_h \tag{8-8}$$

42. The design weights are constant within strata but generally differ across strata. If the sample is proportionally allocated to the strata, meaning that $n_h = n(N_h/N)$, then Equation (8-8) yields $d_{hk} = N/n$ for all strata h and units k, and the design turns out to be self-weighting. As it was already noted for systematic sampling, the fact that the design weights have the same expression for a stratified sampling design with proportional allocation and EPSEM selection within strata as for an unstratified SRSWOR of the same size does not mean that the two designs are equivalent. For instance, if the strata are such that the values of the interest variable y are on average more similar within strata than between strata, then stratification improves precision as compared to a SRSWOR of the same size n.

8.2.1.3. Probability proportional to size sampling

43. Assume a numeric variable x exists in the sampling frame whose values are known and strictly positive for all the population units, that is $x_k > 0$ for all $k \in U$. A probability proportional to size (PPS) design can use variable x as a Measure of Size (MOS) to select a sample of fixed size n such that the inclusion probability of each population unit k is proportional to its MOS x_k. The formal expression of the inclusion probabilities in PPS sampling is:

$$\pi_k = n\frac{x_k}{X} \tag{8-9}$$

where X is the population total of the MOS variable x, $X = \sum_U x_k$. In practice, Equation (8-9) can result in $\pi_k > 1$ for some units characterized by large values of the MOS variable (i.e. when $x_k > X/n$). Suppose this happens for a subset of m units out of the n desired ones. In this case, the inclusion probabilities of those m units are set to 1, whereas the inclusion probabilities of the remaining units are re-calculated through Equation (8-9), using $n' = n - m$ instead of n and setting X to the updated sum of the MOS variable over the restricted sampling frame that excludes the first m units. If probabilities greater than 1 are generated again, the above steps are iterated until $\pi_k \leq 1$ holds true for all the units. Units with inclusion probability 1, and thus design weight 1, are often called "self-representing" units.

44. The design weights in a PPS sampling design that uses variable x as MOS can formally be expressed as follows:

$$d_k = \frac{X}{n\, x_k} \tag{8-10}$$

In single stage sampling designs, PPS selection produces perfect estimates of the population total of the MOS variable, X. In fact, from Equation (8-10), one has: $\hat{X}_{\text{HT}} = \sum_s d_k x_k = \sum_s (X/n) = X$, regardless of the selected sample s. The same would hold true for $\hat{Y}_{\text{HT}}$, if y and x were strictly proportional over the population, $y_k = \text{constant} \times x_k$ for all $k \in U$. For this reason, PPS designs are expected to improve estimation efficiency when the interest variable y and the MOS variable x are correlated. This is often the case of enterprise surveys, where the size of the enterprise in terms of number of employees is typically

used as MOS, and many interest variables, such as value added or investments, are known to positively correlate with the size of the enterprise.

45. As expanded upon in the next section, PPS sampling also plays an important role in household surveys, which often use multistage stratified cluster sampling designs. There, the primary sampling units are typically selected with PPS, using the frame variable "number of households" as MOS. If the second-stage selection of households is properly designed (e.g. by applying sampling rates that are approximately inversely proportional to the MOS of the primary sampling units), this approach has two desirable advantages. From a statistical standpoint, it may lead to a household sample that is approximately self-weighting within strata, thus partially compensating for the efficiency loss induced by clustering and multistage selection. From a fieldwork standpoint, by making the household sample size approximately constant within primary sampling units, it may result in roughly equal interviewer workloads.

8.2.1.4. Multistage stratified cluster sampling in household surveys

46. Most household surveys conducted via face-to-face interviews employ a multistage stratified cluster sampling design. The primary sampling units (PSU), which are randomly selected in the first stage, are usually nonoverlapping geographical areas that form a partition of the target territory, often Census enumeration areas (EA). The secondary sampling units (SSU), which are randomly selected in the second stage within the sampled PSUs, are households. If all household members are surveyed, the design has only two stages. If, within each sampled household, a random subset of eligible household members is surveyed (e.g. one randomly selected woman of reproductive age), then the design has three stages, and the individuals are sometimes referred to as tertiary sampling units (TSU).

47. Some household surveys combine two-stage and three-stage designs. In these cases, a main questionnaire collects data from all household members, while a specialized questionnaire is administered to a random subset of eligible members selected within each household. This approach results in two distinct samples, one from the main questionnaire and one from the specialized questionnaire, each with its own sampling design and target population. Consequently, as outlined in Section 8.1.7, such surveys produce two separate sets of individual weights, with the appropriate weight for estimation determined by the variables involved in the analysis.

Household weights

48. In multistage stratified cluster sampling designs, the calculation of the household-level and individual-level design weights must factor in the inclusion probabilities of clusters selected at each subsequent sampling stage (i.e. PSUs, SSUs, and TSUs). This calculation is greatly simplified because the selection processes of PSUs within strata, households within PSUs, and individuals within households are all stochastically independent. For instance, the inclusion probability of household j, belonging to PSU i, sampled within stratum h, can be expressed as follows:

$$\pi_{hij} = \pi_{hi}^{\mathrm{PSU}} \times \pi_{j\,|\,hi}^{\mathrm{SSU}} \tag{8-11}$$

where π_{hi}^{PSU} is the inclusion probability of PSU i within stratum h, and $\pi_{j\,|\,hi}^{\mathrm{SSU}}$ is the inclusion probability of household j conditional on the first stage selection of PSU i within stratum h.

49. Typically, PSUs are selected independently within strata with PPS and the total number of households resulting from the sampling frame is used as MOS variable. Therefore, from Equation (8-9), π_{hi}^{PSU} can be expressed as follows:

$$\pi_{hi}^{\text{PSU}} = n_h \frac{M_{hi}}{M_h} \tag{8-12}$$

where n_h is the number of PSUs allocated to stratum h, M_{hi} is the total number of households in PSU i of stratum h (i.e. the MOS of the sampled PSU), and M_h is the total number of households across all PSUs in stratum h (i.e. the total MOS of stratum h).

50. At the second stage, households are randomly selected within each sampled PSU. In countries with more advanced statistical infrastructure, the household lists that serve as second-stage sampling frame may come from recent Census files, administrative data, or centralized statistical registers. In low- and middle-income countries (LMICs), where such sources may be outdated or unavailable, a full enumeration exercise is often conducted in each sampled PSUs ahead of the survey, providing fresh lists of households to be used as second-stage sampling frames.

51. The households are typically selected within each sampled PSU with equal inclusion probabilities, often by systematic sampling. Moreover, to achieve a nearly self-weighting sample at stratum-level and optimize fieldwork balance, a fixed number of households, call it m, is selected within each PSU. This allows expressing $\pi_{j\,|\,hi}^{\text{SSU}}$ as follows:

$$\pi_{j\,|\,hi}^{\text{SSU}} = \frac{m}{M_{hi}^*} \tag{8-13}$$

where M_{hi}^* is the total number of households in sampled PSU i of stratum h as resulting from the second-stage list frame. Note that M_{hi}^* may differ from the corresponding MOS value, M_{hi}, reported on the PSU frame. For instance, the enumeration exercise may reveal that $M_{hi}^* > M_{hi}$ because of population growth occurred from the time the PSU frame was constructed.

52. From Equations (8-11)–(8-13), the household-level inclusion probabilities turn out to be:

$$\pi_{hij} = n_h \frac{m}{M_h} \times \frac{M_{hi}}{M_{hi}^*} \tag{8-14}$$

Accordingly, the design weight d_{hij} of household j, belonging to PSU i, sampled within stratum h is:

$$d_{hij} = \frac{M_h}{m\, n_h} \times \frac{M_{hi}^*}{M_{hi}} \tag{8-15}$$

In the absence of nonsampling errors, the design weights in Equation (8-15) produce unbiased estimates of totals, Y, that describe the household population. For instance, the sum over the household sample s of the household-level design weights is an unbiased estimate of the total number of households in the target population, denoted by M: $\widehat{M}_{\text{HT}} = \sum_s d_{hij}$.

53. When the MOS values from the sampling frame are close to the household listing figures, $M_{hi} \cong M_{hi}^*$, an approximately self-weighting sample within each stratum is obtained: $d_{hik} \cong M_h/(m\, n_h)$. In other words, the household inclusion probabilities no longer depend on the PSU and are constant within strata. This design can be made approximately self-weighting even across strata if the n PSUs are allocated proportionally to the household population of the strata. In fact, by setting $n_h = n(M_h/M)$, one obtains: $d_{hik} \cong M/(n\, m)$. These design weights have the usual structure of EPSEM designs, namely the ratio between the total number of households in the population (i.e. M) and total number of households selected in the multistage sample (i.e. $n \times m$).

Individual weights

54. Regarding the design weights of individuals, a distinction must be drawn based on whether the household survey collects data on all household members or on a random subset of them. In the first case, the survey has only two stages of sampling. Individuals belonging to the sampled households are selected with certainty. Therefore, the inclusion probability of each individual equals the inclusion probability of the household to which he/she belongs. For these designs, the inclusion probability of the generic individual k belonging to household j in PSU i within stratum h is:

$$\pi_{hijk} = \pi_{hij} = n_h \frac{m}{M_h} \times \frac{M_{hi}}{M^*_{hi}} \tag{8-16}$$

Accordingly, the design weights of the individuals are:

$$d_{hijk} = d_{hij} = \frac{M_h}{m\, n_h} \times \frac{M^*_{hi}}{M_{hi}} \tag{8-17}$$

As shown in Equation (8-17), individual design weights are constant within households, and each individual has the same design weight as his/her household. Weights with these properties are called "integrated household-individual weights". In the absence of nonsampling errors, the design weights in Equation (8-17) produce unbiased estimates of totals, Y, that describe the individual population. For instance, the sum over the individual sample s of the individual-level design weights is an unbiased estimate of the target population size in terms of individuals, denoted by N: $\widehat{N}_{\mathrm{HT}} = \sum_s d_{hijk}$.

55. In the second case, when the household survey collects data only on a random subset of eligible household members, the survey has three stages of sampling, and individuals play the role of tertiary sampling units (TSU). In this case, because of the third stage of selection, the inclusion probabilities of individuals differ from those of the households to which they belong:

$$\pi_{hijk} = \pi^{\mathrm{PSU}}_{hi} \times \pi^{\mathrm{SSU}}_{j\,|\,hi} \times \pi^{\mathrm{TSU}}_{k\,|\,hij} = \pi_{hij} \times \pi^{\mathrm{TSU}}_{k\,|\,hij} \tag{8-18}$$

56. In Equation (8-18), $\pi^{\mathrm{TSU}}_{k\,|\,hij}$ denotes the inclusion probability of individual k conditional on the second stage selection of household j in PSU i within stratum h. This conditional probability depends on the way individuals are sampled within households. If, for instance, one adult person is randomly selected for interview within each sampled household with equal probability, then $\pi^{\mathrm{TSU}}_{k\,|\,hij}$ is simply the reciprocal of the number of adults in household j, call it a_j, namely $\pi^{\mathrm{TSU}}_{k\,|\,hij} = 1/\,a_j$. Thus, the inclusion probabilities of individuals would be:

$$\pi_{hijk} = \frac{\pi_{hij}}{a_j} = n_h \frac{m}{M_h} \times \frac{M_{hi}}{M^*_{hi}} \times \frac{1}{a_j} \tag{8-19}$$

Accordingly, the design weights of the randomly selected adult individuals would be:

$$d_{hijk} = d_{hij} \times a_j = \frac{M_h}{m\, n_h} \times \frac{M^*_{hi}}{M_{hi}} \times a_j \tag{8-20}$$

Under this three-stage design, the individual weights in Equation (8-20) are not the same as the household weights. The two sets of weights are not integrated. Moreover, in this case, the sum over the individual sample s of the individual-level design weights is an unbiased estimate of the adult population size, denoted by A, not of the size of the general population N: $\hat{A}_{\mathrm{HT}} = \sum_s d_{hijk}$.

57. Note that the logic of the probability calculation illustrated above for the selection of one person among the eligible household members is also applicable when the selection of households is not performed via a two-stage process (e.g. when households are sampled directly from a list of telephone numbers).

58. Some household surveys use multiple questionnaires administered to different respondent samples. Typically, a main questionnaire is given to all household members, while a specialized questionnaire is administered to a random subset of eligible members selected within each household. For instance, in a Demographic and Health Survey (DHS), a specialized questionnaire may collect reproductive health information from a randomly selected woman of reproductive age in each sampled household. The respondents to the main and specialized questionnaires form two logically distinct samples of individual-level observations, each selected with a different sampling design and representing a different target population (in the DHS example, all individuals versus only women of reproductive age). Consequently, as outlined in Section 8.1.7, the survey will yield two separate sets of individual weights. Weights for respondents to the main questionnaire will follow Equation (8-17), whereas weights for those responding to the specialized questionnaire will be calculated along the lines of Equations (8-18)–(8-20). Notably, individuals who respond to both questionnaires will receive two distinct weights, with the applicable weight depending on the variables involved in the analysis. For instance, in the DHS scenario, any analysis focusing on variables derived from the specialized questionnaire, such as estimating the proportion of women of reproductive age using contraceptive methods by age class, will require the specialized weights.

Weights for other analysis units linked to households

59. As outlined in the Introduction, the interviewed households can sometimes be used as a vehicle to collect data on units from linked populations for which a sampling frame does not exist. This approach is known as "indirect sampling" (Lavallée, 2007). For instance, in LMICs, a household survey may include a module on agriculture to collect data at the plot level. Similarly, data on non-farm enterprises (NFE) operated by the households may be collected to study the informal sector. The data collected on these linked units, such as plots and NFEs, can provide valuable contextual information for drawing inferences about households and individuals. However, there are situations where the analysis focuses directly on these linked units. For instance, a researcher might aim to estimate the number of NFEs operated by households in the country and their average monthly sales. In such cases, weights must be attached to the analysis units to produce unbiased estimates. Since no sampling frame exists for these units, their inclusion probabilities cannot be directly computed, and Equation (8-4) is not applicable. Nonetheless, the "generalized weight share method" (GWSM) (Deville and Lavallée, 2006) can be applied to derive the weights of the analysis units from the weights of the households. Provided the links connecting the analysis units and the households are known, and every analysis unit in the target population is linked to at least one household, the GWSM ensures unbiased estimates of the analysis population under indirect sampling.

60. In the NFE example, the GWSM would calculate the weight d_e^{NFE} of the generic NFE e observed by the survey as a linear combination of the weights of the *sampled* households that operate the NFE. Each household weight in the combination would be divided by the *total* number of households that operate that NFE:

$$d_e^{NFE} = \sum_s d_k \frac{L_{ek}}{L_e} \tag{8-21}$$

where the link indicator L_{ek} is 1 if household k operates NFE e and 0 otherwise, and $L_e = \sum_U L_{ek}$ is the sum over the household population U (not the household sample s) of the links of NFE e, namely the *total* number of households (sampled or not) that operate the NFE. Note that the value of L_e is generally

unknown, so a specific question must be included in the NFE module of the questionnaire to obtain it from the operating household.

61. Intuitively, NFEs that are co-operated by a larger number of households L_e have a larger probability to be observed, as they could be indirectly sampled via a larger number of households. This explains why the GWSM downscales their weight by a factor $1/L_e$, as shown in Equation (8-21). In a scenario where NFEs co-operated by multiple households do not exist, Equation (8-21) would produce NFE weights that are equal to the household weights.

8.2.2. Weighting adjustments

62. While design weights serve as the bedrock for survey estimation, further adjustments to design weights are often warranted either to improve estimate efficiency or, more crucially, to provide protection against biases emanating from nonsampling errors. Weight adjustments are implemented according to a clearly specified sequence of steps. Each adjustment step multiplies the weights it receives in input by an adjustment factor. For instance, in a real-world survey, the statistician can adjust the design weights to counteract nonresponse bias and then calibrate the nonresponse adjusted weights to mitigate undercoverage bias and/or improve estimation efficiency. This sequence can be symbolized as follows:

$$d_k = 1/\pi_k \xrightarrow[\text{Adjustment}]{\text{Nonresponse}} w_k^{NR} = a_k^{NR} d_k \xrightarrow{\text{Calibration}} w_k^{CAL} = a_k^{CAL} w_k^{NR} \tag{8-22}$$

where the multiplicative factors a_k^{NR} and a_k^{CAL} are the nonresponse and calibration adjustment factors, respectively. While this example reflects the typical core weighting adjustments implemented in most sample surveys, there often can be other steps that are necessary given the context of the survey. However, the principle of a sequential series of adjustments holds no matter how many different adjustments are required. In what follows, the most common weighting adjustments are described. While not an exhaustive list, the adjustments discussed here will cover the vast majority of circumstances.

8.2.3. Adjustment for unknown eligibility

63. A sampling frame can sometimes include units that do not belong to the survey's target population. For example, in a household survey conducted by telephone, the sampling frame may include telephone numbers of businesses. These ineligible units may contaminate the selected sample and contribute to what is known as overcoverage error of the survey. When the eligibility status of all the sampled units can be determined, either during or after data collection, the ineligible units can be screened out of the sample. In such cases, bias can be avoided, and the overcoverage error's negative effect is limited to a loss of precision in estimates due to reduced data.

64. However, there are situations where it is impossible to ascertain the eligibility status of some sample units, particularly if they are nonrespondents. For instance, in the telephone survey scenario, it can happen that no one answers despite multiple call attempts, making it impossible to know whether the nonresponding unit was a household (eligible) or a business (ineligible). When a sample is known to be contaminated by ineligible units, the survey statistician must assume that some of the units with unknown eligibility may be ineligible. However, since some of these units may also be eligible, simply excluding all units with unknown eligibility from the sample could lead to biased estimates. Therefore, to protect the survey from bias, the weights of units that are known to be eligible must be adjusted to account for the units with unknown eligibility. Only after performing this adjustment can the statistical analysis be restricted to the units known to be eligible.

65. The unknown eligibility adjustment is usually very simple (United Nations, 2008). Basically, it amounts to inflating the weights of the units whose eligibility status is known in such a way that they also account for the sum of the weights of the units with unknown eligibility. Call s_E, s_I, and s_U the subsets of units known to be eligible, known to be ineligible, and whose eligibility status is unknown in the sample s, so that $s = s_E \cup s_I \cup s_U$. Then, the unknown eligibility adjusted weights, call them w_k^{UE}, can be expressed as follows:

$$w_k^{UE} = a_k^{UE} d_k = \left(\frac{\sum_s d_k}{\sum_{s_E \cup s_I} d_k} \right) d_k \quad k \in s_E \cup s_I \tag{8-23}$$

By construction, the adjusted weights w_k^{UE} of the known eligibility subset sum-up to the sum of the weights over the whole sample, $\sum_{s_E \cup s_I} w_k^{UE} = \sum_s d_k$. Moreover, with little algebra, it can be seen that Equation (8-23) implies the following expression for the adjusted weights of the eligible subset:

$$w_k^{UE} = a_k^{UE} d_k = \left(\frac{\sum_{s_E} d_k + \varepsilon \sum_{s_U} d_k}{\sum_{s_E} d_k} \right) d_k \quad k \in s_E \tag{8-24}$$

where $\varepsilon = \sum_{s_E} d_k / \sum_{s_E \cup s_I} d_k$ is the estimate of the frame eligibility rate based on the known eligibility subset. Equation (8-24) is easy to interpret: the weights of the eligible units are inflated in such a way that their sum now also accounts for the sum of the weights of the units expected to be eligible among those whose eligibility is unknown. In fact: $\sum_{s_E} w_k^{UE} = \sum_{s_E} d_k + \varepsilon \sum_{s_U} d_k$.

8.2.4. Adjustment for nonresponse

66. As described in Section 8.1, nonresponse is an ever-present source of nonsampling error in surveys with the potential to introduce bias into survey estimates. Such biases can be counteracted through implementation of adjustments to survey weights. Given the pervasive nature of nonresponse, statisticians have devoted a great deal of attention towards the development of nonresponse adjustment methods. Many prevailing adjustment methods attempt to explicitly model the response propensity of sampled units (both responding and nonresponding) and anchor adjustments to these modeled response propensities. The response propensity modeling (RPM) approaches can vary substantially in complexity and effectiveness which is often a direct function of the auxiliary information that is available for sampled units that can be applied to model response behavior. In addition to RPM approaches, calibration (the subject of Section 8.2.5) can also be an effective approach to address nonresponse bias without explicitly relying on modeling response propensity.

67. In real-world surveys, survey variables are almost never fully observed for all n units belonging to the selected random sample s due to the nonresponse phenomenon. Information can only be collected on $m < n$ units belonging to a subset r of the planned sample s, i.e. $r \subset s$. The response set (or "respondent sample") r is a random set and can be thought as the outcome of two distinct phases: (1) the *sample selection phase* where a random sample s is selected with n units and known inclusion probabilities π_k and, subsequently, (2) the *response phase* where each unit k in s is included in r with a specified response probability (or "response propensity") p_k. While the sampling phase is controlled by the statistician, the response phase is not and thus response probabilities p_k are unknown in practice. Only the ultimate response outcome is observed (whether the unit responded or not) but the more nuanced theoretical probabilities of response for each unit are unobservable. If response probabilities were known and positive, design weights could be easily modified to perfectly deal with nonresponse by substituting planned-sample inclusion probabilities π_k with response-set inclusion probabilities ϕ_k, which are simply the product of the inclusion and response probabilities:

$$\phi_k = \Pr(k \in r) = \Pr(k \in s)\Pr(k \in r|s) = \pi_k p_k \tag{8-25}$$

This would lead to the two-phase extension of the HT estimator in Equation (8-2) (Särndal and Lundström, 2005):

$$\hat{\mathrm{Y}}_{2\mathrm{P}} = \sum_r \frac{1}{\phi_k} y_k = \sum_r \frac{d_k}{p_k} y_k \tag{8-26}$$

68. The two-phase HT estimator $\hat{\mathrm{Y}}_{2\mathrm{P}}$ would be unbiased but unfortunately it is impossible to construct since p_k is unknown. However, $\hat{\mathrm{Y}}_{2\mathrm{P}}$ can be approximated by estimating response probabilities $\hat{p}_k$ through a (explicit or implicit) statistical model, yielding the following nonresponse adjusted empirical estimator:

$$\hat{\mathrm{Y}} = \sum_r \frac{d_k}{\hat{p}_k} y_k = \sum_r w_k^{NR} y_k \tag{8-27}$$

where:

$$w_k^{NR} = d_k/\hat{p}_k \quad k \in r \tag{8-28}$$

is the nonresponse adjusted weight of respondent unit k. Since estimated response probabilities are less or equal to 1, all the design weights either stay the same or receive an upward adjustment. This is in line with intuition, as the nonresponse adjusted weights w_k^{NR} need to compensate for the loss of sample units caused by nonresponse (recall that $r \subset s$). Note that, although in Equation (8-28) the nonresponse adjustment factor is applied to the design weights d_k, in concrete application it could be applied to weights that already incorporate previous adjustments, like the unknown eligibility weights w_k^{UE} of Equation (8-24). Valliant et al. (2018, Chapter 13.5.1) provide further details on the interplay between unknown eligibility and nonresponse adjustments.

69. Formulating the statistical model to estimate $\hat{p}_k$ forms the basis for the most prominent nonresponse weighting adjustment methods through RPM. There are two main classes of RPM approaches that are widely implemented: (1) response homogeneity groups (RHG) or weighting classes and (2) logistic RPM. Both of these approaches are presented below. The RHG model is simpler and technically easier to fit (and is in fact a special case of the logistic model). In both cases, the response propensity model uses as a dependent variable a binary response indicator, which is 1 for respondent units and 0 otherwise:

$$R_k = \begin{cases} 1 \text{ if } k \in r \\ 0 \text{ if } k \in s - r \end{cases} \tag{8-29}$$

70. The unit response probability p_k is modeled as an explicit function $f(\cdot)$ of a set of explanatory variables $\boldsymbol{z}_k = (z_{k1}, \ldots, z_{kp})$ and associated parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$:

$$p_k = \mathrm{prob}(R_k = 1) = f(\boldsymbol{z}_k \boldsymbol{\beta}) \tag{8-30}$$

The model is then fitted to the survey data, yielding estimates of the model parameters $\hat{\boldsymbol{\beta}}$, which finally lead to the needed estimated response probabilities of the respondent units:

$$\hat{p}_k = f(\boldsymbol{z}_k \hat{\boldsymbol{\beta}}) \quad k \in r \tag{8-31}$$

71. When using the RPM approach, the degree of success of the nonresponse-adjusted weights in mitigating nonresponse bias crucially depends on the ability of the adopted model to provide good estimates $\hat{p}_k$ of the unknown true response probabilities p_k. In other words, the RPM weighted estimators of

population totals in Equation (8-27) are model-dependent and their bias depends on the degree of validity of the model. In practice, no response model could ever be fully valid, even if all the possible explanatory variables $\boldsymbol{z}_k$ are observed and available. What is more, many real-world surveys do not have at their disposal a rich set of explanatory variables $\boldsymbol{z}_k$ to use in estimating the model. A very important requirement of the RPM approach is that all explanatory variables $\boldsymbol{z}_k$ must be known for *both respondents and nonrespondents*. This can be a serious constraint to implementing the RPM approach since quite often there is limited or no information available on nonrespondents. Panel surveys can be seen as a lucky exception, as nonrespondent values $\boldsymbol{z}_k$ can often be imputed based on observations from previous waves.

72. In addition to having explanatory variables $\boldsymbol{z}_k$ to feed into the model, the relevance of variables that are included in the model is also a critical factor affecting the validity of the RPM and, in turn, the effectiveness of nonresponse adjustments emanating from the model. In order to be effective, explanatory variables $\boldsymbol{z}_k$ included in the RPM should be correlated with the pattern of and propensity for non-response and, ideally though not critically, correlated with the outcome variables of interest for the survey.

8.2.4.1. Response homogeneity groups

73. The first and simplest application of the RPM approach is through formation of what are called response homogeneity groups (Särndal et al., 1992) or response weighting classes (Little, 1986). The idea behind this approach is to identify nonoverlapping groups within the population (and that can also be identified in the sample) that are expected to exhibit the same or similar response propensity and then implement a simple ratio adjustment to the weights that is constant for all sample members within the group. The crucial and at times complicated task in the RHG approach is the formulation of the groups themselves. There are several common approaches ranging from simplistic (e.g., treating sampling strata as RHGs) to complex (e.g., through use of classification algorithms) which are described below.

74. The underlying assumption of the RHG model is that the population consists of non-overlapping subpopulations (the "groups") such that:

- All the units within each group respond with the same probability,
- Different groups may have different response probabilities, and
- Response/nonresponse outcomes are independent across units.

75. Once the groups themselves are identified, response probabilities $\hat{p}_k$ can be estimated in a very straightforward way. If the RHG model holds, unbiased estimates of response probabilities are given by the observed group-level response rates. Assume g RHGs were formed. Introducing symbols for the subsets of the planned sample s and respondent subsample r belonging to the j-th RHG, s_j and r_j, as well as for their respective sizes, n_j and m_j, then it can be written:

$$\hat{p}_k = \frac{m_j}{n_j} \quad \forall k \in r_j \;\; j = 1, \ldots, g \tag{8-32}$$

It follows then that the nonresponse adjusted weights will simply be:

$$w_k^{NR} = d_k / \hat{p}_k = \frac{n_j}{m_j} d_k \quad \forall k \in r_j \;\; j = 1, \ldots, g \tag{8-33}$$

76. A legitimate alternative way to estimate response probabilities under the RHG model is to use weighted response rates instead of crude, unweighted ones. Instead of taking the ratio of raw counts m_j and n_j as in Equation (8-32), the weighted response rates approach takes the ratio of the sum of design weights

across units in r_j and s_j. Following this approach, the $\hat{p}_k$ would be interpreted as estimates of *population* response rates within the groups rather than *sample* response rates. This leads to:

$$\hat{p}_k = \frac{\sum_{r_j} d_i}{\sum_{s_j} d_i} \quad \forall k \in r_j \;\; j = 1, \; \ldots, \; g \tag{8-34}$$

so that the nonresponse adjusted weights read:

$$w_k^{NR} = d_k / \hat{p}_k = d_k \frac{\sum_{s_j} d_i}{\sum_{r_j} d_i} \quad \forall k \in r_j \;\; j = 1, \; \ldots, \; g \tag{8-35}$$

In concrete applications, when it comes to bias mitigation, both the weighted and the unweighted variants often perform similarly (Valliant et al., 2018).

77. The modeling effort for the RHG model lies in the problem of forming groups that result in nearly constant response probability within each group. This is a highly nontrivial task, as the response mechanism can arguably be very complex. In addition, variables that are expected to play an important role in explaining nonresponse can unfortunately be unavailable in practice (especially because their values are not known on nonrespondent units). Therefore, RHGs are often postulated based on the few variables that are available for both respondents and nonrespondents. For instance, RHGs are sometimes simply identified with sampling strata even though, on substantive grounds, there is no compelling reason to believe that response probabilities are truly constant within these "compromise" RHGs. Despite its limited nature, it is always advisable to try to adjust the weights at a minimum using sampling strata as RHGs. In fact, this will likely decrease nonresponse bias to some extent, albeit perhaps not decisively.

78. Beyond simply taking the sampling strata as RHGs, the groups can be formed following a variety of different approaches. The critical requirements for forming the groups are (1) that the groups are nonoverlapping and (2) that they completely partition the sample. If only a handful of variables are available for all sampled units, then the groups could simply be formed by crossing multiple categorical variables (e.g., sex and age groups). If a greater set of variables is available, then more advanced methods could be applied to construct the groups. One possibility that is discussed in the next sub-section and proposed by Little (1986) is to estimate a more refined response propensity model and use the predicted $\hat{p}_k$ not as a direct adjustment factor but instead as a means to form RHGs according to quantiles of the $\hat{p}_k$ distribution (an approach sometimes referred to as "propensity stratification"). Another common method is to apply classification algorithms to form the RHGs. Classification algorithms essentially are mathematical algorithms that analyze data across multiple variables and try to identify patterns to segment the data into classes along those patterns. There is an abundance of different classification algorithms which can be applied (too many to holistically enumerate here) but some examples (as listed by Valliant et al., 2018) are classification and regression trees (CART; Breiman et al., 1984), support vector machines (Vapnik, 1995), chi squared automatic interaction detection (CHAID; Kass, 1980), and random forests (Breiman, 2001). While it is beyond the scope of this handbook to provide details on how to implement these algorithms, Valliant and Dever (2018) as well as Valliant et al. (2018) provide further details and practical applications to forming RHG using classification algorithms.

8.2.4.2. Logistic response propensity model

79. Another widely used nonresponse adjustment method is the logistic response probability model which attempts to explicitly model response propensity applying multiple covariates ($\boldsymbol{z}_k$) in a binary choice framework. While the RHG model uses just a single categorical explanatory variable to perform nonresponse adjustments (i.e. the identifier of the groups), the logistic RPM can use many explanatory

variables, both categorical and numeric, to explicitly model and predict $\hat{p}_k$. The unit response probability p_k is modeled as a binary choice, typically applying a logistic[5] function of the explanatory variables $\boldsymbol{z}_k = (z_{k1}, \ldots, z_{kp})$ and the associated parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$:

$$\text{logit}(p_k) = \log\left(\frac{p_k}{1-p_k}\right) = \boldsymbol{z}_k\boldsymbol{\beta} \tag{8-36}$$

Thus, the estimated response probabilities of the respondent units are:

$$\hat{p}_k = \frac{e^{\boldsymbol{z}_k\hat{\boldsymbol{\beta}}}}{1+e^{\boldsymbol{z}_k\hat{\boldsymbol{\beta}}}} \quad k \in r \tag{8-37}$$

which implies that the nonresponse adjusted weights under the logistic RPM model are:

$$w_k^{NR} = d_k/\hat{p}_k = d_k\frac{1+e^{\boldsymbol{z}_k\hat{\boldsymbol{\beta}}}}{e^{\boldsymbol{z}_k\hat{\boldsymbol{\beta}}}} \quad k \in r \tag{8-38}$$

The ability of the logistic RPM to accommodate multiple explanatory variables (including interaction terms between them) can result in better estimates of response probabilities and improved mitigation of nonresponse bias. While this added complexity of the logistic RPM makes it technically more complicated to implement than RHG, the modeling and subsequent calculation of $\hat{p}_k$ based on the model are easily implemented in most common statistical software.

80. There are also some important factors to be wary of with the logistic RPM approach. One concern is the potential for extreme values for the nonresponse adjustment factors $1/\hat{p}_k$ which can in turn result in unstable estimates. The more complex the logit model, the larger will be the variability of the adjustment factors and hence of the adjusted weights w_k^{NR}. For comparison, the RHG model, which is a special case of the logistic RPM with just one categorical predictor with g classes, generates only g different adjustment factors while the logit RPM generates unit-specific adjustment factors. To avoid the generation of extreme nonresponse adjusted weights that can result in unstable estimates, two techniques (collectively referred to as propensity stratification) inspired by the RHG approach are often used in practice. The first technique is implemented as follows:

[1] Compute quantiles (e.g. quintiles or deciles) of the distribution of the estimated response probabilities $\hat{p}_k$,

[2] Use the quantiles to split the respondent sample into RHGs,

[3] Estimate group-level response probabilities using group-level averages of the estimated $\hat{p}_k$, and then,

[4] Obtain nonresponse adjusted weights by dividing the design weight by the average estimated response probability within the RHG.

81. This first technique leverages the estimated logistic model in two ways: to define the RHGs (using quantiles of the $\hat{p}_k$) and to estimate the corresponding group-level response probabilities (using averages of the $\hat{p}_k$). The second technique integrates the RPM and RHG approaches even more extensively. It only

[5] While the logistic function is the predominant link function used for explicit RPMs, it is also possible to apply a probit or complementary log-log model. These three models generally provide similar results and thus the choice of which model to apply is largely arbitrary.

uses the fitted logistic model to smartly split the respondent sample into RHGs, but no longer to estimate the response probabilities within them. In this way, some robustness is gained against misspecifications of the RPM model. This second technique is implemented as follows:

[1] Compute quantiles of the distribution of the estimated response probabilities $\hat{p}_k$ (as before),

[2] Use the quantiles to split the respondent sample into RHGs (also as before),

[3] Estimate group-level response probabilities as in an ordinary RHG approach, using weighted or unweighted group-level response rates (instead of the average estimated response probabilities from the RPM model),

[4] Obtain nonresponse adjusted weights by dividing the design weight by the group-level response rates estimated in step iii.

8.2.4.3. Non-response adjustments for multistage samples

82. When samples are drawn according to a multistage design, it is possible for nonresponse to occur among units sampled at each stage. For example, in a traditional two-stage household survey design with EAs as PSUs and households as SSUs, it is possible that in addition to households who do not respond, there may be some EAs that do not "respond". Perhaps some EAs were inaccessible by interviewers as a result of some climatic event. Similarly, when information on some of the household members of a respondent household cannot be collected, nonresponse occurs at the individual level. In such instances, nonresponse could be handled separately (and sequentially) at each stage of the design. While the nature of the nonresponse and information available for implementing adjustments will differ across the different stages, the principles outlined above for implementing the adjustments are always applicable.

8.2.5. Calibration adjustment

83. Calibration (Deville and Särndal, 1992) is a methodology to improve the quality of survey estimates by using, in a systematic and rigorous way, auxiliary information on the target population that is available from external sources. As symbolically shown in Equation (8-22), household surveys often apply a calibration adjustment to the nonresponse adjusted weights to gain an additional layer of protection against biases (e.g. undercoverage bias and/or residual nonresponse bias), improve the precision of survey estimates, or ensure alignment of survey estimates with population totals from trusted external sources. In addition, calibration is also increasingly being used to address nonresponse, serving as an alternative to the RPM approach (Särndal and Lundström, 2005). In such cases, the calibration process uses as input the design weights (potentially adjusted for unknown eligibility). While the calibration methodology can be tailored to achieve a variety of inferential objectives, the degree of success in meeting these objectives depends on the nature and richness of the available auxiliary information.

84. As outlined in Section 8.1.6, calibration weights are obtained by minimally adjusting the input weights so that survey estimates exactly match population totals known from external sources. The survey variables for which population totals are available are called auxiliary variables. In household surveys, population totals are typically counts of households and individuals, either for the target population as a whole or by domains. Possible examples are the total number of households in the country, the number of households by province and rural/urban status, the total number of individuals in the country, the number of individuals by region, sex, and age classes, the total number of individuals by household size classes. Note that calibration can simultaneously address multiple auxiliary variables, so that a single calibration step would generally be sufficient to exactly match all the mentioned population totals. In addition to counts, calibration can also exploit numeric auxiliary variables whose population totals are amounts, for instance

the total yearly household income of the country, or the total monthly household expenditure for food and nonfood items. However, in the context of household surveys, reliable sources for such numeric totals are rarely available. This is at odds with enterprise surveys, where calibration processes routinely use numeric auxiliary variables whose population totals are available from a business register.

8.2.5.1. The calibration problem

85. From a mathematical standpoint, calibration weights are the solution of a constrained optimization problem. The calibration problem can be formalized as follows[6]:

$$\begin{cases} \text{argmin}_{\{w_k\}} \sum_r G(w_k, d_k) & \text{subject to:} \\ \sum_r w_k \boldsymbol{x}_k = \boldsymbol{X} \\ \text{L} \leq \dfrac{w_k}{d_k} \leq \text{U} \end{cases} \tag{8-39}$$

where the sums run over the units of the eligible respondent sample, $k \in r$, and:

- w_k is the sought set of calibration weights, which must be obtained by solving Problem (8-39).
- d_k is the set of input weights to the calibration adjustment step.
- $G(w_k, d_k)$ is a distance function that measures how close the set of calibration weights w_k is to the set of input weights d_k. Different choices of the distance function $G(\cdot,\ \cdot)$ are possible.
- $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{kp})$ is a vector of p auxiliary variables observed on each eligible respondent unit k.
- $\boldsymbol{X} = (X_1, \ldots, X_p)$ is the vector of known population totals of the p auxiliary variables, namely $X_j = \sum_U x_{kj}$ for $j = 1, \ldots, p$.
- $[\text{L}, \text{U}]$ are a lower and an upper bound that restrict the acceptable range of the calibration adjustment factors, the so-called g-weights, $g_k = w_k/d_k$.

86. The first two equations of Problem (8-39) constitute integral parts of the calibration methodology, whereas the third equation is optional. The first equation formalizes the objective of finding calibration weights that are as close as possible to the input weights. The second equation defines "calibration constraints", whose purpose is to ensure that the calibrated estimates will exactly match the population totals. The third equation imposes range restrictions on the size of the calibration adjustment factors. Although optional, such range restrictions are commonly applied to prevent the generation of negative or exceedingly high calibration weights.

87. Note that, if calibration is used to address nonresponse, the d_k in Problem (8-39) must be understood as design weights (potentially adjusted for unknown eligibility). Otherwise, the d_k must be understood as

[6] For the sake of space, the multiplicative unit-specific factors $1/q_k$ that appear in the distance minimization objective of (Deville, Särndal, 1992) are not discussed. They are optional and seldom used in household surveys, though they are often useful in enterprise surveys.

survey weights that already incorporate a previous nonresponse adjustment. To accommodate both meanings, in what follows the d_k will be generically referred to as "initial" weights.

8.2.5.2. Calibration estimators

88. As was the case for the design weights, also the calibration weights are "universal". Once a solution to Problem (8-39) is found, the same calibration weights w_k can be used to calculate survey estimates for any interest variable y. The calibration estimator of the population total of variable y is defined as:

$$\hat{Y}_{\text{CAL}} = \sum_r w_k \, y_k \tag{8-40}$$

89. By construction, the calibration estimator produces estimates of the totals of the auxiliary variables that exactly match the input known population totals, $\hat{\boldsymbol{X}}_{CAL} = \boldsymbol{X}$, that is:

$$\hat{X}_{j\,\text{CAL}} = \sum_r w_k \, x_{kj} = X_j = \sum_U x_{kj} \qquad j = 1, \ldots, p \tag{8-41}$$

90. The solution of Problem (8-39):

$$w_k = g_k \, d_k \qquad k \in r \tag{8-42}$$

can be expressed in analytic closed form only if (i) $G(\cdot, \cdot)$ is the "chi-square distance" function $G(w_k, d_k) = (w_k - d_k)^2/(2d_k)$ and (ii) no range restrictions are imposed on the g-weights (namely, L $\rightarrow -\infty$ and U $\rightarrow \infty$). In this case, the calibration estimator in Equation (8-40) is identical to the Generalized Regression Estimator (GREG), and calibration weights can legitimately assume negative values (Särndal et al., 1992).

91. Under any other settings, the calibration weights w_k can only be obtained by solving Problem (8-39) via iterative numerical optimization methods. Specialized software packages exist that can compute calibration weights in large-scale surveys even under complex specifications of auxiliary variables and calibration constraints. Some examples, as listed by Valliant et al. (2018, page 370), are the *survey* package (Lumley, 2004 and 2011) and *ReGenesees* (Zardetto, 2015) in R, the *WTADJUST* and *WTADJX* procedures in *SUDAAN* (RTI International, 2012), the *svycal* function (Valliant and Dever, 2018) and the *ipfraking* package (Kolenikov, 2014) in Stata.

92. Based on Equations (8-40) and (8-5), calibration estimators and Horvitz-Thompson estimators may appear deceptively similar. However, they are fundamentally different. Horvitz-Thompson estimators are linear functions of the design weights, whereas calibration estimators depend nonlinearly[7] on the d_k. Consequently, Horvitz-Thompson estimators are unbiased, whereas calibration estimators are only asymptotically unbiased, meaning their bias approaches zero as the sample size becomes large. In addition, the nonlinearity of calibration estimators makes the problem of estimating their sampling variance more complex than it would be for Horvitz-Thompson estimators (Deville and Särndal (1992), Demnati and Rao (2004)).

93. In practice, the bias attributable to calibration adjustments of the weights is always negligible in large-scale surveys, provided the population totals $\boldsymbol{X}$ are true and error-free. In other words, under this condition, even if the success of calibration in reducing nonresponse or undercoverage bias depends on the

[7] This is because, due to the structure of Problem (8-39), the calibration weight w_k of each sample unit k does not only depend on d_k but also on the initial weights d_j of all the other sample units $j \neq k$.

auxiliary variables, calibration is at least guaranteed *not to introduce* additional bias in the weighting system of the survey. This critically important property of calibration estimators hinges upon the requirement that calibration weights w_k must be as close as possible to the initial weights d_k (Särndal, 2007). Therefore, the minimization of the distance function $G(\cdot,\ \cdot)$ plays a crucial role in Problem (8-39) and should not be overlooked. Heuristic or subjective modifications of the weights that disregard this minimization may result in introducing bias even if the adjusted weights happen to satisfy all the specified calibration constraints.

8.2.5.3. Popular examples of calibration estimators

94. Many popular estimators traditionally known under other names are special cases of calibration estimators. The post-stratified estimator (see, e.g., Holt and Smith (1979), Deville and Särndal (1992)) and the raking estimator (see, e.g., Deming and Stephan (1940), Deville et al. (1993)) are two of the most well-known examples.

The post-stratified estimator

95. To implement this estimator, the required population totals are the population counts across a partition of the population, referred to as "post-strata". Like sampling strata, post-strata must be mutually exclusive and collectively exhaustive. However, unlike stratification variables, the post-stratification variables do not need to be known for all units in the sampling frame. In this approach, the identifiers of the post-strata serve as auxiliary variables for calibration and thus only need to be known for the respondent units. Although post-stratification is applied after data collection, it provides many of the benefits of stratification in improving estimate precision.

96. A key feature of the post-stratified estimator is that the calibration weights w_k can be expressed in a simple analytic form as a function of the initial weights d_k:

$$w_k = d_k \frac{N_p}{\sum_{r_p} d_i} \quad \forall k \in r_p \;\; p = 1,\ \ldots,\ P \tag{8-43}$$

where index p identifies the post-stratum, r_p is the respondent subsample belonging to the post-stratum, and N_p is the known population count for post-stratum p. The working mechanism of the calibration adjustment is very clear: units belonging to post-strata whose size would be "underestimated" using the initial weights have their weights inflated, and vice versa. This way, the calibration weights result in perfect estimates of the size of the post-strata:

$$\widehat{N}_p^{\mathrm{CAL}} = \sum_{r_p} w_k = N_p \quad p = 1,\ \ldots,\ P \tag{8-44}$$

The raking estimator

97. The raking estimator is another widely used calibration method. It adjusts the survey weights to simultaneously match the known population counts for two or more categorical auxiliary variables, such as sex and age classes. Importantly, raking does not require knowledge of the population counts for the cross-classified cells of these variables (e.g. the intersections of sex and age classes). If such detailed cross-classified totals were available, post-stratification could be used instead of raking.

98. The strength of raking lies in its ability to simultaneously match the known totals for multiple auxiliary variables. This makes the approach especially useful when the joint distribution of the variables in the population is unknown but matching any of the individual margins alone would not be sufficient. However, unlike post-stratification, raking does not yield a simple closed-form solution for the calibration weights. Instead, the weights are determined through an iterative numerical process (e.g. the Iterative

Proportional Fitting algorithm), which continues until the adjusted weights produce estimates that match the known population totals for each auxiliary variable.

8.2.5.4. Integrated household-individual calibration weights

99. Many household surveys collect data from all household members, resulting in integrated household-individual design weights, where all individuals share the same weight as the household to which they belong. As noted in Section 8.1.7, preserving this integration after weight adjustments ensures an internally consistent system of household-level and individual-level estimates. It can also improve statistical efficiency by reducing variability in individual-level weights.

100. However, when individual-level population totals are used, standard calibration methods do not automatically produce identical calibration weights across members of the same household. Achieving integrated household-individual calibration weights requires specialized algorithms. These algorithms were first proposed in Lemaître and Dufour (1987) and Heldal (1992). Both methods follow a similar approach. First, the calibration model matrix formed by staking the auxiliary vectors $\boldsymbol{x}_k = (x_{k1}, \dots, x_{kp})$ of the m respondent individuals, $k \in r$, is aggregated at household-level. Then, the calibration task is performed on this aggregated dataset. Finally, the obtained household-level calibration weights are re-expanded to the individual-level, by attaching to each individual the calibration weight of the household the individual belongs to. The only difference between the techniques lies in how the initial weights and auxiliary variables are treated: the former approach sums the initial weights and averages the auxiliary variables across individuals within each household, whereas the latter averages the initial weights and sums the auxiliary variables.

8.2.5.5. Desirable properties of the auxiliary variables

101. Thanks to its versatility, calibration can be employed to achieve various inferential objectives. Historically, the methodology was developed to enhance the precision of survey estimates under ideal conditions, namely in the absence of nonsampling errors. While reducing the sampling variance of estimators remains a desirable goal, modern household surveys increasingly apply calibration to address biases originating from nonsampling errors, such as undercoverage or nonresponse bias. The book by Särndal and Lundström (2005) lays the methodological foundations for this approach, discussing calibration techniques aimed at improving estimates under nonresponse and frame imperfections.

102. The success of calibration in reducing the sampling variance of key survey estimators and/or mitigating bias crucially depends on the nature and richness of the available auxiliary information. Below, the desirable properties of auxiliary variables to achieve specific inferential goals are concisely described.

Improvement of estimate precision

103. As shown in Equation (8-41), calibration ensures perfect estimates of the totals of the auxiliary variables, $\widehat{\boldsymbol{X}}_{CAL} = \boldsymbol{X}$, where $\boldsymbol{X}$ represents the known, error-free population totals available from external sources. In other words, the calibration estimators $\widehat{\boldsymbol{X}}_{CAL}$ no longer depend on the realized respondent sample r and thus have zero sampling variance. Intuitively, by reducing the sampling variance of the estimated totals of the auxiliary variables to zero, calibration is expected to decrease the sampling variance of any interest variables that are correlated with the auxiliary variables. More formally, for large samples, the sampling variance of the calibration estimator $\hat{Y}_{\mathrm{CAL}}$ tends to be smaller than that of the Horvitz-Thompson estimator $\hat{Y}_{\mathrm{HT}}$ by a factor of $1 - \mathrm{R}^2$, where R^2 is the coefficient of determination of the census regression of the variable y on the p auxiliary variables $x_1, \dots, x_p$. Consequently, the stronger the correlation between the interest variable and the auxiliary variables, the greater the gain in efficiency from calibration.

104. In Section 8.2.5, it was observed that the known totals used to calibrate household surveys are often limited to counts of households and individuals by domains determined by territorial and socio-demographic variables (e.g. region, rural/urban status, household size classes, sex, age classes). Although simple, such known totals and auxiliary variables are nonetheless expected to be beneficial in reducing the sampling variance of key survey indicators (e.g., the unemployment rate for a Labor Force Survey). This is because territorial and socio-demographic auxiliary variables typically do correlate with the interest variables involved in those indicators (in the LFS example, the labour status of the individuals).

Mitigation of nonresponse bias

105. In recent years, the calibration approach to nonresponse has been steadily gaining ground. Contrary to the response propensity model (RPM) method discussed in Section 8.2.4, calibration does not require the auxiliary variables to be known for nonrespondents. This feature makes calibration appealing, since little is generally known on nonrespondents, especially for cross-sectional surveys.

106. The book by Särndal and Lundström (2005) identifies and motivates key criteria for selecting powerful auxiliary variables when calibrating for nonresponse. To put it simply, reduction of nonresponse bias can be achieved by calibration if the auxiliary variables:

- Are correlated with the response propensity Criterion (a)
- Are correlated with the interest variables, Criterion (b)
- Identify the main estimation domains of the survey. Criterion (c)

These three criteria form a hierarchy, and, ideally, powerful auxiliary variables should simultaneously meet all of them. However, each property is beneficial on its own.

107. Criterion (a) is essential to effectively reduce nonresponse bias across *all potential interest variables*. This requirement is akin to the conditions needed for a successful response propensity model (RPM), as discussed in Section 8.2.4. A fundamental difference is that calibration, unlike the RPM approach, does not require the user to formulate an explicit model to guide the nonresponse adjustment. Instead, one can think as if the calibration adjustment factors, g_k, of Equation (8-42) implicitly estimate the reciprocals of the response probabilities, $1/\hat{p}_k$, of Equation (8-28).

108. Criterion (b) is the fundamental condition for reducing sampling variance when using calibration estimators under ideal conditions. In real-world calibration scenarios, fulfilling this requirement reduces nonresponse bias for *specific interest variables* that are directly correlated with the auxiliary variables.

109. Together, criteria (a) and (b) provide what is known as "double protection" against nonresponse bias (see Kott and Liao, (2012), and references therein). Double protection means that bias reduction for a variable y is achieved if either criterion (a) or (b) holds true for the used auxiliary variables, even if both are not satisfied simultaneously.

110. Criterion (c) plays a role in reducing both bias and sampling variance of domain estimates.

111. To better illustrate how criteria (a), (b), and (c) work in practice, consider the following example. Suppose a Labor Force Survey (LFS) was conducted in a LMIC and experienced significantly different response rates across rural and urban areas. Response rates were higher in rural areas, where unemployment rates have historically been lower. Additionally, labour market status (employed, unemployed, inactive) is known to correlate, albeit mildly, with demographic factors such as age, sex, and education level. The LFS aims to produce official estimates of labour market indicators at both national and regional levels. In this

scenario, nonresponse could introduce bias, such as underestimating the unemployment rate, unless the survey weights are suitably adjusted. If calibration is used to adjust the weights for nonresponse, good auxiliary variables could include the following:

- For criterion (a): Rural/urban status. Since response rates differ markedly across rural and urban areas, this variable directly correlates with the likelihood of responding.
- For criterion (b): Age, sex, and education level. These variables are linked to labor market status, the key interest variable of the LFS, and thus would help reduce both bias and variance in the estimates.
- For criterion (c): Region. The LFS will report estimates at the regional level, so ensuring that regional population totals are incorporated in the calibration helps ensure accurate domain-level estimates.

112. While it would be ideal to use all these auxiliary variables, in practice, the necessary population totals may not be available for some. For example, there might not be a reliable and up-to-date source on the population distribution by education level. This would force the statistician to calibrate the survey using only the remaining four auxiliary variables: rural/urban status, age, sex, and region. The example shows that, in addition to identifying good auxiliary variables, also finding reliable external sources for their population totals is a crucial step to complete a successful calibration task.

Mitigation of undercoverage bias

113. When units not covered by the sampling frame differ systematically from those that are covered, undercoverage can lead to biased estimates unless the survey weights are appropriately adjusted. Calibration can mitigate undercoverage bias, if (i) the known population totals come from higher-quality external sources with better coverage and currency than the frame, (ii) the auxiliary variables can account for some of the differences between covered and non-covered units, and (iii) the auxiliary variables correlate with the variables of interest. The rationale is that, by removing any undercoverage bias from the estimates of the auxiliary variables' totals, calibration weights enable the covered sample units to better represent the unobserved, non-covered portion of the population.

114. For example, consider a mobile phone survey conducted in a LMIC and aimed at estimating the proportion of households with bank accounts. Since non-mobile phone owners (often poorer households) are excluded from the frame, wealthier households will be overrepresented. This biases the unadjusted estimate of bank account ownership upwards. However, calibration using recent census data on region, rural-urban status, household size, and education and labour status of the household head, could mitigate this bias, as these variables help predict both mobile phone and bank account ownership.

8.2.5.6. Sources of population totals

115. As discussed in Section 8.2.5.2, calibration estimators are unbiased for large-scale surveys when the population totals used in the constraints are true and error-free. However, no source of population totals is ever perfectly accurate. Even successful censuses inevitably suffer from small coverage issues. This makes the quality of the source of auxiliary information a key factor for the success of calibration.

116. For household surveys, possible sources of population totals that can be used in calibration tasks include:

[A1] The sampling frame used to select the sample
[A2] The most recent population and housing census

[A3] Administrative archives or statistical registers maintained by the country's NSO
[B1] A national master sample maintained by the country's NSO
[B2] A recent, high-quality sample survey whose target population coincides with or includes the target population of the survey at hand
[B3] Population projections coming from a high-quality demographic projection exercise
[B4] The survey itself, for nonresponse calibration tasks where auxiliary variables are available for both respondents and nonrespondents

117. The listed sources naturally fall into two different types: A and B. Type A sources may not be perfect in practice, but, at least in principle, if they were so, they would produce true and error-free population totals, $\boldsymbol{X}$. Conversely, type B sources cannot ever produce ideal population totals, not even in principle. This is because type B sources inherently produce estimates, $\widetilde{\boldsymbol{X}}$, rather than known population totals, $\boldsymbol{X}$. Both B1 and B2 produce estimates that are affected by sampling variability, whereas estimates B3 embody the many model uncertainties of the underlying demographic forecasting technique (e.g. variants or simulation errors). When estimated population totals, $\widetilde{\boldsymbol{X}}$, coming from type B sources are used to set the calibration constraints in Problem (8-39), their uncertainty necessarily transmits to the resulting calibration estimators. Nonetheless, as long as these estimated population totals are unbiased, the resulting calibration estimators will still be useful in reducing bias, and the only price for using type B sources will be a loss of estimation efficiency. Although papers exist (e.g. (Berger et al., 2009) and (Dever, Valliant, 2016)) addressing the problem of estimating the sampling variance of calibration estimators that use estimated population totals, the related methods are not yet implemented in most software packages.

118. The preferable choice among the available sources of calibration totals can only be identified on a case-by-case basis. For instance, when calibration must be used to gain protection against undercoverage bias, source A1 is obviously ruled out. In LMICs, census data may sometimes be severely outdated, making source A2 problematic. Sources A3 are often only accessible in high-income countries. The recency and quality of master samples (B1) and population projections (B3), which are often available in LMICs, deserve to be scrutinized as well. Although sources B2 have a significant potential to address the scarcity of type A sources in LMICs, they have traditionally been used primarily in high-income countries. For instance, the Canadian and Italian LFS surveys, which use rotating panel designs, enhance the precision of labor market estimates by using estimated calibration totals derived from the overlapping portion of the LFS sample in previous waves. Finally, the use of source B4 when calibrating for nonresponse is discussed in depth in the book by Särndal and Lundström (2005). Just to mention an example, the RHG adjusted weights in Equation (8-35) can be seen as calibration weights where the population totals are the estimated counts of population units in the groups, $\sum_{s_j} d_i$, as resulting from the *planned* sample, namely *both* respondents and nonrespondents.

8.2.6. Cumulative impact of weighting adjustments on variability

119. All of the adjustments described in the preceding sections are in furtherance of the goal of counteracting the potential for bias in survey estimates. While necessary, these adjustments can have another less desirable consequence through increased variability of the survey weights which, in turn, can result in deteriorated precision of survey estimates. This is most impactful when weighting adjustments generate extreme outlier weights for some units. Furthermore, adjustments in some circumstances can lead to "uncomfortable" weights that are less than one or even negative. While the impact on estimate variance is typically not as pronounced for these uncomfortable weights, they can be difficult for data users to digest, particularly from the common interpretation that the magnitude of the weights reflects "the number of population units the sample unit represents".

120. There are trimming methods available to deal with both extreme and uncomfortable weights, however they are not without their risks and should not be applied without first conducting an assessment to establish the need for as well as potential impact of trimming. One means to conduct such an assessment is to examine the unequal weighting effect (UWE) developed by Kish (1992). The UWE is a useful generalized metric to assess the impact of variability of the weights on estimate precision. It builds upon the notion of the design effect (described in Chapter 5, Section 5.9.2) with the UWE serving as another component, in addition to the clustering effect, that contributes to the relative inefficiency of a realized complex sample design when compared to simple random sampling:

$$DEFF = UWE \times DEFF_{clustering} \tag{8-45}$$

where $DEFF_{clustering}$ is the design effect due to clustering and UWE is the design effect due to unequal weighting. Kish defines UWE as follows:

$$UWE = \frac{n\sum_k w_k^2}{(\sum_k w_k)^2} = 1 + cv_{w_k}^2 \tag{8-46}$$

The UWE measure can be interpreted analogously to $DEFF_{clustering}$ where a value of 1 indicates that there is no inflation of variance as a result of weighting (i.e., in the case of equal weighting) while the higher the value, the more inflation of variance due to unequal weights. The UWE can serve as a very useful metric to assess the impact of a particular weighting adjustment as well as the cumulative impact of all weighting adjustments.

121. To diminish the inflated estimate variance that results from highly variable weights there are two classes of prevailing techniques: (1) those that prevent excessive weights when implementing adjustments and (2) those that "trim" or bound the final weights after all adjustments are implemented.

122. Perhaps the most effective way to counteract outlier or uncomfortable weights is simply to tweak the methods of implementing weighting adjustments to prevent them from happening in the first place. For example, when implementing nonresponse adjustments using logistic response propensity model, instead of using the unit-level values of $1/\hat{p}_k$ as adjustment factors it is advisable to form response classes according to the distribution of $\hat{p}_k$ and use the average $1/\hat{p}_k$ within the class (as described in Section 8.2.4.2). This will diminish the potential for outlier values of $\hat{p}_k$ resulting in outlier weights. For calibration adjustments, specifying lower and upper bound values $[\mathrm{L}, \mathrm{U}]$ of the calibration adjustment factors g_k can eliminate the potential for uncomfortable weights and reduce the potential for outlier weights generated through calibration (as described in Section 8.2.5.1). While these are two prominent examples, the general approach is to be mindful of potential ways to constrain adjustments to prevent generation of troublesome weights without compromising substantially the positive impacts of the adjustments. As emphasized above, such decisions must first be informed by an assessment of the impact of the adjustment on the weights and a determination that the adjustment has resulted in uncomfortable or outlier weights that could be limited through modification of the adjustment approach itself.

8.2.6.1. Trimming adjustment

123. While the first approach is certainly preferred to prevent outlier weights, if after conducting all weighting adjustments outlier weights are still detected, then there are a variety of techniques available to address them through "trimming" or winsorizing. While these techniques do succeed in reducing the variability of the weights, they also can introduce bias into the survey estimates, particularly when trimming is implemented heuristically or haphazardly. Of course, the hope is that the bias induced by trimming the

weights (if any) turns out to be much smaller than the resulting gain in estimation precision (i.e. reduction of estimators' standard errors).

124. Most weight trimming methods start by (1) identifying a certain threshold above (in the case of excessively large weights) or below (for excessively small weights) which weights are considered outliers, (2) reassigning outlier weights to the relevant threshold and (3) redistributing the differential value of the trimmed weights across the other units not subjected to trimming. This basic approach to weight trimming is summarized by Haziza and Beaumont (2017) as:

$$w_k^{TRIM} = \begin{cases} w_l & \text{if } w_k \leq w_l \\ w_u & \text{if } w_k \geq w_u \\ \gamma w_k & \text{if } w_l < w_k < w_u \end{cases} \quad (8\text{-}47)$$

where w_k^{TRIM} is the version of the weight after trimming, w_l and w_u are the lower and upper thresholds which identify outliers, and γ is a rescaling factor which guarantees that trimming does not alter the population size estimate based on the respondent sample r:

$$\sum_r w_k^{TRIM} = \sum_r w_k \quad (8\text{-}48)$$

125. While this framework summarizes the generic approach, there are many variations in its implementation. Establishing the thresholds of w_l and w_u is the first critical step. They should not be established heuristically or haphazardly but rather through an informed process based on the distribution of weights. A common yet often undesirable practice is to take a certain percentile (e.g., the 1st or 99th) as the thresholds for identifying and trimming outlier weights. While in some circumstances such thresholds could be effective, there is a risk that they will miss some outlier weights that happen to fall within these thresholds or (more concerningly) will trim weights that, while small or large, are not excessively so and carry important information about the inclusion probability of the unit. Potter and Zheng (2015) identify three additional methods to establish these thresholds, though they are neither exhaustive nor definitive. Unfortunately, there is no optimal and universal approach for establishing these thresholds.

126. After the thresholds are identified, the extreme weights are typically set to the relevant threshold values. However, only adjusting the value of these outlier weights will break the correspondence between the weights and the estimated population size or the calibration constraints (if calibration was performed). The last (and sometimes neglected) step in the weight trimming approach is to apply a further adjustment to the weights that were not identified as outliers which ensures that these correspondences are maintained. A constant adjustment factor γ in Equation (8-47) above is enough to preserve the estimated population size as dictated by Equation (8-48). This desirable outcome is, for example, guaranteed if weights are trimmed using the *survey* package in R. However, more sophisticated approaches are needed to simultaneously preserve multiple calibration totals. For instance, if the weights were post-stratified, the trimming adjustment should ideally maintain the sum of weights within each post-stratum. This implies that a different adjustment factor γ_p for each post-stratum p would be needed. For more general and complex calibration constraints, specialized software such as *SUDAAN*, *ipfraking*, and *ReGenesees* allow users to trim the weights to a specific interval $[w_l, w_u]$ while simultaneously preserving all the calibration totals. A commonly used technique to achieve this result is the GEM (Generalized Exponential Model) proposed by Folsom and Singh (2000).

127. At what point should trimming be applied? While it may be tempting to implement weight trimming after each adjustment step, this practice is not recommended and rather trimming should be implemented

(if deemed necessary) *after all weight adjustments have been implemented*. As emphasized above, trimming can introduce bias and thus implementing trimming at multiple points in the weight construction process invites the possibility of introducing more and more bias at each trimming step while also destroying some of the meaningful information contained in the untrimmed weights. Furthermore, outlier weights generated at one step might be automatically mitigated or counteracted at a subsequent step and therefore prevent the need for trimming this weight at the end of the process. Given the risks associated with trimming, it is optimal to (1) implement trimming sparingly and in an informed way, avoiding heuristic approaches and (2) only trim using the final weights that result after all adjustment steps have been completed.

8.2.7. Weight calculation diagnostics

128. Given the critically important role that weights play in providing reliable estimates from probability sample surveys, it is important to undertake the weight calculation process with much care and humility. Mistakes in the calculation process or undesirable properties that result from the adjustments could lead to inefficient weights or, more detrimentally, weights that produce biased estimates. This sub-section provides a set of diagnostics that can and should be implemented to identify potential issues and ensure the integrity of the weights.

129. The previous sections have demonstrated the sequential process of weight calculation often involving many adjustments that inherit intermediate weights from the previous step. Any potential flaws emanating from a particular adjustment will continue to reverberate through later steps and even potentially be masked from detection. It is therefore important to perform checks and diagnostics on the adjusted weights generated at each step before moving to a subsequent.

130. There are two broad classes of diagnostics which could be considered: (1) definitive checks to confirm that the theoretical properties of the weights are satisfied and internally consistent and (2) diagnostics to monitor the quality of the weights produced at each stage.

8.2.7.1. Definitive checks

131. The purpose of "definitive checks" is to make sure the produced weights possess the theoretical properties expected of them. Therefore, the failure of these checks unequivocally indicates flaws in the weight calculation process, requiring corrective measures.

132. There are several definitive checks on the weights that can be performed at different steps of the weight calculation process. The most critical step at which to implement such checks is when calculating the design weights. These checks must be performed by the person responsible for selecting the sample and calculating the design weights, as he/she has complete access to both the sampling frame and the sample. If, after the survey is completed, another person inherits the design weights and is tasked with calculating the final survey weights, he/she will no longer be able to perform all the necessary checks, having access only to the respondent sample $r \subset s$.

133. Of course, the most basic check that should be performed is to confirm that all units in s have a nonmissing design weight that is greater than or equal to 1. Another slightly more sophisticated but nonetheless necessary check is to confirm the consistency between the sum of the design weights and the corresponding totals from the frame. The relevant correspondence to check is dependent on the sample selection procedure adopted. For example, in household surveys, under a two-stage design with equal probability of selection of PSUs, the PSU-level design weights must sum up to the total number of PSUs in the frame. Under a two-stage design that utilizes PPS for PSU selection, the sum of PSU design weights no longer needs to exactly match the number of PSUs in the frame but is instead expected to approximate it, with any large deviations requiring investigation. In this scenario, however, an alternative definitive

check exists, namely that the PSU design weights multiplied by the number of households in the PSU must sum up to the total number of households according to the frame.

134. If information on the total number of households is not available from the frame (a potential restriction particularly for those that inherit design weights and did not perform the selection themselves), it could be possible to compare the implied totals from the design weights with external data sources that provide an estimate of the total number of households. While precise correspondence with the external benchmark is not necessarily expected, this is another case where large deviations could be indicative of a problem with the design weights and prompt further investigation.

135. In addition to performing this diagnostic on the design weights themselves, it is also highly recommended to confirm approximate correspondence between the benchmark value for the total number of households and the sum of adjusted weights *after each adjustment step*. Large discrepancies may indicate errors in the implementation of the adjustments or that the adjustments were inadequate to address the biases affecting the survey.

136. There are other definitive checks that apply to specific steps in the weight construction process. For example, when implementing a nonresponse adjustment via RPM (or RHG) there must be no instance where the adjusted weight is lower than the input weight. The adjustment should always result in an inflation of weights for the responding sample. At the calibration step, another definitive check to perform is to confirm that after calibration the adjusted weights result in estimates that exactly match the population totals used as calibration constraints. While most software packages perform this check automatically, independently verifying this alignment remains a valuable practice.

8.2.7.2. Diagnostics for quality monitoring

137. In addition to the definitive checks, a variety of other diagnostics can be performed to monitor and assess the quality and suitability of the weights produced at each adjustment step. Several important diagnostics are described here, although many more exist beyond those mentioned. The purpose of these diagnostics is not to check for errors or inconsistencies (which are largely handled by the definitive checks described above) but to identify potential inefficiencies in the weights and explore ways to improve their efficiency.

138. One of the most basic diagnostics that should be performed is to compare the distribution of the weights before and after applying an adjustment. This can easily be achieved through generating a simple scatter plot and visually inspecting the plot. The idea is to try to (1) gain a better understanding of the direction and consistency of the weight adjustment and (2) to identify cases where there is an excessively large adjustment worthy of scrutiny. A further variation that could be implemented is instead to plot the input weights against the adjustment factors themselves.

139. As highlighted in Section 8.2.6, weighting adjustments often increase the variance of the weights, which can reduce the precision of estimates. Therefore, it is crucial to monitor how each adjustment step affects weight variability. Kish's UWE (introduced in Chapter 5 and discussed in Section 8.2.6) is a valuable metric for this purpose because it provides interpretable values that can be directly compared to understand the impact on the variance of the weights. To evaluate an adjustment's impact, calculate the UWE for both the input and adjusted weights, and compare the two values. The absolute level of UWE is less important than the change in UWE before and after the adjustment. An increase in UWE indicates more weight variability due to the adjustment. While increased variance is generally undesirable, it is fully justified if the weight adjustment effectively reduces biases, leading to more reliable estimates. Therefore, a higher post-adjustment UWE should not cause immediate concern. However, if an adjustment dramatically raises

the UWE, it would be wise to review and possibly refine the adjustment method to achieve the goal of reducing bias while keeping the increase in weight variance to a minimum.

140. In addition to assessing the impact of adjustments on the variability of the weights, it would also be beneficial to analyze the impact of weight adjustments on key indicators from the survey. A simple way to achieve this could be by computing weighted estimates of the key indicator(s) with the input and adjusted weights and comparing the magnitudes. If a particular adjustment results in a substantial change in the indicator(s), this could be a sign to investigate further and see if alternative approaches to implementing the adjustment yield a less dramatic impact. However, this comparison should be made with care since if the weighting adjustment is implemented to counteract a significant source of bias, then perhaps a large change in the indicator could be expected and even desired.

141. In addition to simply comparing estimates using the two weights, it could also be informative to compute the product of each weight and the variable of interest and display the two products in a scatter plot. With a visual inspection, this could identify troublesome cases where an adjusted weight w_k^*, though not excessively large on its own, combines with a large but seemingly acceptable value of the interest variable y_k^*, to produce an exceptionally large product $w_k^* y_k^*$. Such influential values $w_k^* y_k^*$ could lead to unstable estimates and inflated standard errors. It would then be beneficial to investigate the cause for both the relatively large weight and indicator value to see if they are justified or if the weighting adjustment approach might be tweaked to prevent the generation of the influential value $w_k^* y_k^*$.

142. At the calibration step, there are also a number of diagnostics that can be performed. One straightforward check is to compare the weighted population estimates prior to the calibration adjustment with the calibration benchmarks themselves. Large discrepancies between the estimates and benchmarks would (1) indicate that the calibration task is substantial and therefore a nontrivial increase in the variance of the weights might be expected, (2) point to a potential issue emanating from a previous weight adjustment step which has caused this disparity, or (3) bring into question the suitability and compatibility of the calibration benchmarks. This diagnostic step can highlight these issues and prompt further investigation and necessary adjustments.

143. The diagnostics shared in this section provide a flavour of the type of checks that could be performed to ensure that the weights resulting from the entire weight calculation process are as good, accurate, and efficient as possible. Further diagnostics beyond those presented here are certainly available and could be beneficial.

8.3. Specialized topics

144. This section outlines weighting considerations for panel surveys and pooled samples, which do not follow the standard steps and methods detailed in Section 8.2.

8.3.1. Panel surveys

145. In panel surveys, the same sample units are surveyed repeatedly over time, with each survey occasion referred to as a wave. In the simplest setup, a random sample is selected and surveyed at wave 1 and subsequently resurveyed in each following wave, meaning no new random selection occurs after the initial wave. Thanks to their time dimension, panel surveys can support two types of statistical analyses, each requiring a distinct set of weights: cross-sectional and longitudinal (see, e.g., Lynn (Ed.), 2009 and 2021).

146. Cross-sectional analysis of panel data focuses on making inferences about the survey's target population at a specific reference time, namely the time of the panel wave in question. As such, each wave in a panel survey requires its own set of cross-sectional weights. At wave 1, cross-sectional weights are

constructed following the general steps outlined in Section 8.2. However, starting from wave 2, the weighting process no longer begins with design weights, as no further random selection is conducted. Instead, each subsequent panel wave $k > 1$ uses as "initial weights" the "final weights" from the prior wave, that is, the nonresponse-adjusted and calibrated weights of wave $k - 1$. In household panel surveys, these "carryover weights" must then be adjusted to account for changes in household composition over time. Specifically, the weight share method (see, e.g., Kalton and Brick, (1995)) is employed to (i) assign an initial weight to individuals who were not part of the sampled households at wave 1 but joined these households in later waves, and (ii) adjust the household weights accordingly. This adjustment decreases the weight of households that gain new members, as such households could have been included in the sample in multiple ways at wave 1 and therefore have a higher probability to be included in the current wave. After the weight share adjustment, additional nonresponse and calibration adjustments are applied. Two important considerations arise in this process. First, response propensity modeling is often highly effective in panel surveys, as data from earlier waves provide a rich array of explanatory variables known for nonrespondents, as noted in Section 8.2.4. Second, maintaining cross-sectional representativeness over time in panel surveys can be challenging due to attrition, life-cycle dynamics of panel members, and structural changes in the target population. Calibration, which aligns cross-sectional estimates to known population totals at each wave, plays a critical role in mitigating these effects, helping to preserve the cross-sectional representativeness of the survey as the panel ages.

147. Longitudinal analysis in panel surveys focuses on making inferences about changes across waves, such as estimating gross changes (e.g. how many individuals who were in poverty at wave 1 had exited poverty by wave 2) or estimating parameters of longitudinal models (e.g. the survival function of unemployment spells based on individuals who responded to all panel waves). These estimates require linking and jointly analyzing data from multiple waves. Consequently, longitudinal weights are essential to ensure that the linked datasets represent the correct longitudinal populations. For example, the dataset of individuals who responded to both wave 1 and wave 2 should represent the eligible population that existed at both times, something that cannot be achieved by using either of the two wave-specific cross-sectional weights. A practical challenge with longitudinal weights is that their number increases with the survey duration. A panel with K waves can, in theory, represent up to $2^K - K - 1$ distinct longitudinal populations. For long panels, producing this high number of longitudinal weights is thus unfeasible (for example, a ten-wave panel would require over one thousand distinct weights). However, most panel surveys at least produce, at each wave k, the longitudinal weight of those units that responded at all panel waves to date, $1, 2, \ldots, k$. The construction of longitudinal weights does not necessarily follow the steps and methods described in Section 8.2. For instance, to address panel attrition, one can either fit a single response propensity model for the entire longitudinal dataset or a sequence of wave-on-wave response models. Moreover, calibration is typically not applied to longitudinal weights, as totals of auxiliary variables are generally unknown for longitudinal populations. For a more detailed discussion on these and other weighting issues in panel surveys, see Lynn and Watson (2021).

8.3.2. Pooled samples

148. In some cases, multiple independent survey samples $s_1, \ldots, s_p$ may exist that observed the same interest variables y either on the same target population $U = U_1 = \cdots = U_p$ or on a common subpopulation of special analytical interest $\tilde{U} \subseteq U_1 \cap \ldots \cap U_p$. In such cases, researchers may wish to combine ("pool") these independent samples to leverage the larger overall sample size, thereby achieving more precise estimates. When the common subpopulation of analytical interest $\tilde{U}$ is rare, pooling may even be a necessity, as no single survey alone may have captured a sufficiently large sample to provide estimates of satisfactory precision. Furthermore, pooling can serve as a broader design strategy, as in multiple-frame surveys, where

independent samples are selected from distinct but overlapping frames and then combined to estimate characteristics of the target population.

149. Examples of pooled survey samples appear in many practical applications. For instance, long-running household panel surveys occasionally select a fresh "top-up" sample and combine it with the existing panel to increase sample size and better capture recent immigrants. Samples can also be combined over time, as in the case of quarterly Labor Force Surveys, whose quarterly samples are routinely pooled to represent the average population over a survey year and produce annual labor market estimates.

150. Once the independent samples are combined to form the pooled dataset $s_{\text{pool}} = s_1 \cup \ldots \cup s_p$, the original weights of the units $w_k^{(j)}$, where $k \in s_j$ and $j = 1, \ldots, p$, cannot be used as they are, as doing so would lead to incorrect inferential results. This issue is immediately evident when all the samples represent the same population $U = U_1 = \cdots = U_p$. In this case, each separate sample produces unbiased estimates of the population total Y, so using the unadjusted original weights in the pooled dataset would overestimate Y by a factor of p. The problem arises because each unit of the pooled sample could have been included in any of the p samples, but its original weight only reflects the probability of inclusion in the sample where the unit was actually selected. Therefore, to ensure valid analysis of the pooled sample, a "pooling adjustment" factor must be applied to the original weights. This adjustment decreases the original weights to account for the multiple routes of selection in the pooled sample. Following (Chu, Brick, and Kalton, 1999), the adjusted weight for each unit in the pooled sample can be expressed as:

$$w_k^{\text{pool}} = \theta_j w_k^{(j)} \quad \forall k \in s_j \;\; j = 1, \ldots, p \tag{8-49}$$

where the pooling adjustment factors θ_j must be nonnegative and satisfy $\sum_{j=1}^{p} \theta_j = 1$. Using Equation (8-49), the pooled estimator $\hat{Y}_{\text{pool}}$ can be expressed as a weighted average of the individual sample estimators $\hat{Y}_j$, with weights given by the adjustment factors θ_j:

$$\hat{Y}_{\text{pool}} = \sum_{s_{\text{pool}}} w_k^{\text{pool}} y_k = \sum_{j=1}^{p} \sum_{s_j} \theta_j \, w_k^{(j)} y_k = \sum_{j=1}^{p} \theta_j \, \hat{Y}_j \tag{8-50}$$

If the independent samples $s_1, \ldots, s_p$ produce unbiased estimates of the population total when analysed separately, then the above formula clearly guarantees that unbiased estimates will also be obtained from the pooled sample.

151. The adjustment factors in Equation (8-49) can be computed in various ways. The simplest choice would be to set $\theta_j = 1/p$, thereby making $\hat{Y}_{\text{pool}}$ the arithmetic mean of the individual estimators $\hat{Y}_j$. This choice is, however, recommended only when the individual samples are characterized by similar size and estimation efficiency (in fact, $\theta_j = 1/4$ is often used when pooling quarterly LFS samples). Considering the independence of the samples, it is easy to prove that the choice that minimizes the sampling variance of the pooled estimator of the total from Equation (8-50) depends on the sample sizes and the Design Effects of the individual samples (see, e.g., (Lohr, 2021) and references therein):

$$\theta_j = \frac{n_j / \text{Deff}_y^{(j)}}{\sum_{i=1}^{p} n_i / \text{Deff}_y^{(i)}} \qquad j = 1, \ldots, p \tag{8-51}$$

Equation (8-51) highlights that a relatively larger adjustment factor θ_j is assigned to observations from samples with a larger effective sample size $n_j/\text{Deff}_y^{(j)}$.

152. Since Design Effects are specific to each estimator, the pooling adjustment which optimizes the precision of the pooled estimator of the total of variable y, will likely be sub-optimal for other variables. This is inconvenient when the pooled sample must be used for multi-purpose or multivariate statistical analysis. In such cases, the approach of O'Muircheartaigh and Pedlow (2002) offers a natural compromise solution, where variable-specific Design Effects in Equation (8-51) are replaced with the variable-agnostic but survey-specific Unequal Weighting Effects of the individual samples:

$$\theta_j = \frac{n_j/\text{UWE}_j}{\sum_{i=1}^{p} n_i/\text{UWE}_i} \qquad j = 1, \ \ldots, \ p \tag{8-52}$$

Equipped with the pooling-adjusted weights from Equations (8-49) and (8-52), the pooled dataset can then be used to calculate point estimates of any population parameters through standard procedures.

8.4. Emerging approaches: Data integration methods for unbiased estimation from nonprobability samples

153. As discussed in Chapter 1, Section 8.1.3, probability sampling and design-based inference have long served as methodological pillars of official statistics. In recent years, however, there has been a renewed and growing interest in nonprobability surveys (see, e.g., Beaumont (2020)), particularly in high-income countries, where many NSOs have undertaken modernization initiatives to integrate probability surveys with nontraditional data sources.

154. While nontraditional data sources offer advantages in affordability, speed, and scale, deriving accurate and reliable estimates from nonprobability surveys remains a persistent challenge and an active area of research (see, e.g., Zhang (2019), Wu (2022), and references therein). Nonprobability samples are generated through participation mechanisms that are either non-random (i.e. deterministic) or random but unknown and uncontrolled. When randomness is absent or not controlled, the samples are prone to bias and lack sufficient information for analysts to apply effective corrections. For instance, the inclusion probabilities for some population units may be zero, resulting in undercoverage bias (think of the non-online population in a Facebook survey). Even when inclusion probabilities are always greater than zero, they are unequal across units and unknown, making it impossible to adjust for these differences, which leads to selectivity bias (think of an opt-in Web panel that overrepresents people with strong political convictions). Notably, a growing body of theoretical research and empirical evidence suggests that these biases are often exacerbated as the size of the nonprobability sample increases (see, e.g., Meng (2018) and Bradley et al. (2021)). In summary, the fundamental limitation of nonprobability samples is the inability to calculate design weights, which would make the data representative of the target population.

155. In recent years, various "data integration" approaches have been proposed to address the risks of bias in analysing nonprobability samples (see, e.g., Rao (2021) and references therein). Despite their technical differences, these methods share a common principle: combining nonprobability data with information (either at the macro or micro level) from datasets known to be representative, such as census data or probability samples. This integration hinges on auxiliary "bridge" variables x that must:

[1] Fully explain the selectivity mechanism underlying the nonprobability data source, and/or
[2] Effectively predict the variable of interest y in the target population.

156. When these two highly nontrivial assumptions about the bridge variables are met, the integration process allows the nonprobability sample to "borrow representativeness" from the traditional dataset, thereby achieving the desired bias reduction effect.

157. The prevalent data integration approaches for nonprobability samples can be classified into four broad methodologies:

[1] Propensity Modeling (e.g. Chen et al. (2020))
[2] Pseudo-Calibration (e.g. Elliott and Valliant (2017), Rao (2021))
[3] Sample Matching (e.g. Rivers (2007))
[4] Massive Imputation (e.g. Kim et al. (2021))

158. The first two methodologies depend critically on the validity of the first assumption, namely that the bridge variables fully explain the participation mechanism of the nonprobability sample. In contrast, the last two approaches critically rely on the second assumption, namely that the bridge variables adequately predict the variables of interest. The "Doubly Robust" estimation methodology combines these two classes of approaches, achieving unbiased estimates when at least one of the two assumptions holds, even if the other does not (Chen et al., 2020).

159. **Table 8.1** summarises the typical data integration setup. The probability sample A, equipped with survey weights w, is representative of the target population. It includes the bridge variables x but does not contain observations of the variable of interest y. Conversely, the nonprobability sample B lacks survey weights (and thus is prone to bias), includes the bridge variables x, and contains observations of the variable of interest y. The goal is to employ datasets A and B to produce unbiased estimates of the population total of the interest variable $Y = \sum_U y_k$.

Table 8.1. The typical data integration setup to compute estimates from nonprobability samples

Dataset	Source	Sampling	Representativeness	y	x	w
A	Survey	Probability	Guaranteed	○	✓	✓
B	Nontraditional	Nonprobability	Unknown	✓	✓	○

160. The first two approaches, Propensity Modeling and Pseudo-Calibration, employ techniques that are routinely used to address nonresponse and undercoverage bias in probability surveys (see Sections 8.1.5 and 8.1.6). This analogy can nonetheless be misleading, as it obscures a fundamental distinction between the two contexts. In probability surveys, rock-solid design weights exist, and response propensity modeling and calibration are used to *adjust* such weights. By contrast, in nonprobability surveys, design weights are absent, and Propensity Modeling and Pseudo-Calibration must *generate* survey weights from scratch. It should be evident that the second task is formidably harder than the first. Indeed, while the main determinants of nonresponse in probability surveys are well understood, the data-generating mechanism of the nonprobability sample is unknown, making it extremely difficult to identify bridge variables x that can effectively describe its selectivity. Furthermore, even when researchers have a sound hypothesis about the selectivity mechanisms affecting the nonprobability sample B, some of the relevant bridge variables x may not be available in any existing probability sample A.

161. The Propensity Modeling approach combines A and B to create inclusion probabilities for the units of the nonprobability sample B. Specifically, first (i) the indicator R_i of participation in the nonprobability sample ($R_i = 1$ if $i \in A$ and 0 otherwise) is modeled as a function of the bridge variables x, $\text{prob}(R_i = 1 \mid x_i) = \pi_i(x_i)$; then (ii) the model is fitted using x values coming from both A and B and applied to produce estimated inclusion probabilities $\hat{\pi}_k(x_k)$ for the units in the nonprobability sample B; finally (iii) the estimated inclusion probabilities of the enriched dataset B are employed to estimate the total $\hat{Y} = \sum_B y_k/\hat{\pi}_k$. Importantly, Chen et al. (2020) showed that the model fitting step (ii) cannot be performed naively, as in nonresponse adjustments for probability surveys, because this would lead to biased results. A more sophisticated technique is required, which is described in the same paper.

162. The Pseudo-Calibration approach jointly uses datasets A and B to create weights for the units of the nonprobability sample B. First, (i) arbitrary pseudo-weights $\tilde{d}_k$ are assigned to the units of dataset B. Common choices include $\tilde{d}_k = 1$ or $\tilde{d}_k = N/n_B$, where N is the target population size, and n_B is the sample size of B. Then, (ii) dataset A is used to compute unbiased estimates of the population totals of the bridge variables $\hat{X} = \sum_A x_k w_k$, and final weights $\tilde{w}_k$ are generated by calibrating the nonprobability sample B to the estimated population totals $\hat{X}$. Finally, (iii) the pseudo-calibration weights of the enriched dataset B are employed to estimate the total $\hat{Y} = \sum_B y_k \tilde{w}_k$. While both Pseudo-Calibration and Propensity Modeling rely on restrictive assumptions about the bridge variables x, the implementation of the Pseudo-Calibration approach is considerably simpler.

163. Loosely speaking, the last two approaches, Sample Matching and Massive Imputation, leverage the bridge variables to "transport" the interest variable from dataset B to dataset A. More specifically, they (i) exploit the nonprobability sample B to model the relationship $y \sim f(x)$ in the target population; then (ii) apply the resulting model to predict / impute the values of the interest variable $\hat{y}_k = \hat{f}(x_k)$ in the probability sample A; and lastly (iii) use the survey weights w_k of the enriched dataset A to estimate $\hat{Y} = \sum_A \hat{y}_k w_k$. The distinction between the two methods lies primarily in the prediction or imputation technique employed: Massive Imputation typically uses parametric models, for instance a regression model, while Sample Matching relies on non-parametric methods, such as nearest-neighbor imputation.

164. Although the data integration approach provides a modern and promising methodology for deriving unbiased estimates from nonprobability samples, its success depends on restrictive modeling assumptions. These include the completeness of the bridge variables and their ability to explain the variables of interest or the participation mechanism of the nonprobability sample. When these assumptions fail, the methodology cannot safeguard against estimation bias. This is especially concerning because nonprobability samples lack straightforward measures, like the response rate in probability surveys, to gauge the scale of potential bias when assumptions are violated (Meng, 2022). For these reasons, while NSOs are increasingly exploring the use of nonprobability samples for specialized applications, probability sampling remains the dominant paradigm in Official Statistics.

8.5. References

Lavallée P. (2007). Indirect sampling. Springer.

Deville, J., Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. Survey Methodology, 32(2), 165.

Deville, J. C., Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.

Särndal, C. E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. New York: Springer-Verlag.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. Survey methodology, 33(2), 99-119.

Valliant, R., Dever, J. A., Kreuter, F. (2018). Practical tools for designing and weighting survey samples. New York: Springer.

Lumley, T. (2004). Analysis of complex survey samples. Journal of statistical software, 9, 1-19.

Lumley, T. (2011). Complex surveys: a guide to analysis using R. John Wiley & Sons.

Zardetto, D. (2015). ReGenesees: An advanced R system for calibration, estimation and sampling error assessment in complex sample surveys. Journal of Official Statistics, 31(2), 177-203.

RTI International. (2012). SUDAAN Language Manual, Release 11.0. Research Triangle Park, NC.

Valliant, R., Dever, J. A. (2018). Survey weights: a step-by-step guide to calculation. College Station, TX: Stata Press.

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). The Stata Journal, 14(1), 22-59.

Särndal, C. E., Lundström, S. (2005). Estimation in surveys with nonresponse. John Wiley & Sons.

Lemaître, G., Dufour, J. (1987). An integrated method for weighting persons and families. Survey Methodology, 13(2), 199-207.

Heldal, J. (1992). A Method for Calibration of Weights in Sample Surveys. Working Papers from Department for Statistics on Individuals and Households. Methods for Collections and Analysis, N. 3/1992. Oslo, Norway: Central Bureau of Statistics.

Kott, P. S., Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. In Survey Research Methods (Vol. 6, No. 2, pp. 105-111).

Berger, Y. G., Muñoz, J. F., Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals—An application to the extended regression estimator and the regression composite estimator. Computational statistics & data analysis, 53(7), 2596-2604.

Dever, J. A., Valliant, R. (2016). General regression estimation adjusted for undercoverage and estimated control totals. Journal of Survey Statistics and Methodology, 4(3), 289-318.

Kish, L. (1992). Weighting for unequal Pi. Journal of Official Statistics, 8, 183-200.

Lynn, P. (Editor) (2009) Methodology of Longitudinal Surveys. John Wiley & Sons.

Lynn, P. (Editor) (2021) Advances in Longitudinal Survey Methodology. John Wiley & Sons.

Kalton, G., Brick, J. M. (1995). Weighting schemes for household panel surveys. Survey Methodology, 21(2), 33-34.

Lynn, P., Watson, N. (2021). Issues in weighting for longitudinal surveys. Advances in Longitudinal Survey Methodology, 447-468.

Chu, A., Brick, J.M., Kalton, G. (1999). Weights for combining surveys across time or space. Bulletin of the International Statistical Institute, 2, 103-104.

Lohr, S. (2021). Multiple-frame surveys for a multiple-data-source world. Survey Methodology, 47(2), 229-263.

O'Muircheartaigh, C., Pedlow, S. (2002). Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97. In American Statistical Association Proceedings of the Joint Statistical Meetings (pp. 2557-2562). American Statistical Association.

Beaumont, J. F. (2020). Are probability surveys bound to disappear for the production of official statistics? Survey Methodology, 46(1), 1-29.

Zhang, L. C. (2019). On valid descriptive inference from non-probability sample. Statistical Theory and Related Fields, 3(2), 103-113.

Wu, C. (2022). Statistical inference with non-probability survey samples. Survey Methodology, 48(2), 283-311.

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. The Annals of Applied Statistics, 12(2), 685-726.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X. L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. Nature, 600(7890), 695-700.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. Sankhya B, 83(1), 242-272.

Chen, Y., Li, P., Wu, C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532), 2011-2021.

Elliott, M., and Valliant, R. (2017). Inference for nonprobability samples. Statistical Science, 32, 249-264.

Rivers, D. (2007, August). Sampling for web surveys. In Joint Statistical Meetings (Vol. 4). Alexandria, VA: American Statistical Association.

Kim, J. K., Park, S., Chen, Y., Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society Series A: Statistics in Society, 184(3), 941-963.

Meng, X. L. (2022). Comments on "Statistical inference with non-probability survey samples" - Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. Survey Methodology, 48(2), 339-360.

Haziza, D., Beaumont, J. F. (2017). Construction of Weights in Surveys: A Review. Statistical science, 32(2), 206-226.

Potter, F., Zheng, Y. (2015, August). Methods and issues in trimming extreme weights in sample surveys. In Proceedings of the American Statistical Association, Section on Survey Research Methods (pp. 2707-2719). Alexandria, VA: American Statistical Association.

Folsom, R.E., Singh, A.C. (2000) The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 598-603.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). Classification and Regression Trees. New York: Chapman & Hall.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. New York: Springer-Verlag.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means. International Statistical Review, 139-157.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 29(2), 119-127.

Biemer, P. P., Christ, S. L. (2012). Weighting survey data. In International handbook of survey methodology (pp. 317-341). Routledge.

Kalton, G., Flores-Cervantes, I. (2003). Weighting methods. Journal of official statistics, 19(2), 81.

Demnati, A., Rao, J. N. K. (2004). Linearization variance estimators for survey data. Survey Methodology, 30(1), 17-26.

United Nations. Statistical Division. (2008). Designing household survey samples: Practical guidelines (Vol. 98). United Nations Publications.

Horvitz, D. G., Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association, 47(260), 663-685.

Deming, W. E., Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics, 11(4), 427-444.

Deville, J. C., Särndal, C. E., Sautory, O. (1993). Generalized raking procedures in survey sampling. Journal of the American statistical Association, 88(423), 1013-1020.

Holt, D., Smith, T. F. (1979). Post stratification. Journal of the Royal Statistical Society Series A: Statistics in Society, 142(1), 33-46.

8.6. Resources

R survey package
URL: https://cran.r-project.org/package=survey

ReGenesees (R Evolved Generalized Software for Sampling Estimates and Errors in Surveys)
URL: https://diegozardetto.github.io/ReGenesees/

WTADJUST and WTADJX procedures in SUDAAN (add-on to SAS)
URL: https://www.rti.org/impact/sudaanr-statistical-software-analyzing-correlated-data

Stata package ipfraking
URL: https://ideas.repec.org/c/boc/bocode/s458430.html

Stata svycal function (and other svy commands for survey data analysis)
URL: https://www.stata.com/

CHAPTER 9
ANALYSIS

Andrés Gutiérrez (Economic Commission for Latin America and the Caribbean)
Pedro Luis do Nascimento Silva (Sociedade para o Desenvolvimento da Pesquisa Científica)

CHAPTER 9
ANALYSIS

9.1. Introduction

1. A key concern for every agency that produces statistical information is ensuring the *correct* use of the data it provides. This concern is enshrined in the United Nations *Fundamental Principles of Official Statistics,* particularly in the following principles:

- **Principle 3:** To facilitate a correct interpretation of the data, statistical agencies must present information according to scientific standards, including details on the sources, methods, and procedures used.
- **Principle 4:** Statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

2. The advent of the computer revolution, coupled with greater access to computational tools, has led to increased use of statistical data, including household survey data. Sometimes this data is used for mostly *descriptive purposes*, such as estimating population means or obtaining estimates of population frequency distributions. Other times, however, its use is made for *analytical purposes*, involving the testing of hypothesis or the construction of models, when the objective is to draw conclusions that are also applicable to populations other than the one from which the sample was extracted. When using standard statistical software for such analyses, results can be biased or misleading if the complex sampling design is not properly accounted for.

3. Household surveys also play a critical role in tracking progress toward global objectives, such as the **Sustainable Development Goals (SDGs)**. For this purpose, descriptive analyses often include a range of specialized indicators designed to monitor outcomes like access to education, health services, and economic opportunities. These indicators are derived from the survey data and are essential for policymakers and organizations aiming to achieve sustainable development targets.

4. This chapter aims to empower users to analyse household survey data accurately and effectively. It does this by presenting relevant models, methods, and software that enables the data analyst to understand key steps in the data analysis process and to incorporate complex designs into their analyses. It relies on the fundamental concepts of the design-based paradigm for survey design and analysis.

5. What makes household survey data special or challenging for those who intend to analyse them is because they are collected through complex sampling methods that often involve:

- **Stratification:** Dividing the population into comprehensive distinct subgroups before sampling
- **Clustering:** Grouping units and sampling groups rather than units to simplify data collection
- **Unequal probabilities of selection:** Giving units different probabilities of being selected
- **Weighting adjustments:** Correcting for non-response and/or improving precision

6. Standard data analysis methods and software ignore these features, leading to biased estimates of both the target parameters and their associated variances. Here, the impact of simplifications made when using standard data analysis methods and software are analysed, and the necessary adjustments to these methods are presented in order to appropriately incorporate the aspects highlighted above into the analysis.

7. Different readers of the chapter may find some of its parts more useful than others. Here, the contents of each of the sections is described, so that readers may direct their attention to the topics of relevance to

them. Section 9.1 provides guidance on the preparation of a plan for the analysis of the household survey data. Such plans are important both to guide survey development and subsequently any secondary survey data analysis.

8. Section 9.2 provides a short discussion on the fundamental principles of the design-based inference, emphasizing that conclusions taken from probability sample surveys should be based on a pair: the point estimate and it associated margin of error (or any related measure). Section 9.3 begins the journey with the key tools for descriptive analysis: means, ratios, proportions and other typical descriptive parameters.

9. Section 9.4 is devoted to more complex parameters that allow comparisons of the phenomenon of interest between subgroups for continuous and discrete variables. It also presents standard tests to compare means and measure the degree of correlation and association between variables.

10. Section 9.5 focuses on modelling survey outcomes. It starts with a discussion on the role of weighting when estimating regression coefficients, followed by presentation of some proper approaches to estimate complex parameters in linear and logistic regression models. Finally, Section 9.6 presents a summary of ideas and tools for survey data visualization, showing the best practices for creating graphics in a context where sampling imbalance and uncertainty are important.

11. Throughout the chapter, practical examples are provided to illustrate how National Statistics Offices (NSOs) conduct various types of household survey data analysis. These examples provide a useful guide for applying the concepts and methods discussed in real-world contexts. By the end of the chapter, readers will be equipped with the knowledge and tools needed to analyse household survey data effectively while accounting for the complexities of survey designs adopted in such surveys.

9.2. Planning and preparation for analysis

12. Planning the analysis stage of a survey is an essential part of the overall survey planning process. Following the **Generic Statistical Business Process Model (GSBPM),** this step corresponds to the subprocess labelled as *2.1 - Design Outputs* (https://unece.org/statistics/documents/2019/01/standards/gsbpm-v51).

13. At this stage, it is important to distinguish between two groups of users: primary data producers and secondary data users. Proper planning and understanding of the survey design are crucial for both. For primary data producers, creating a comprehensive tabular plan ensures alignment with survey objectives. For secondary users, clear research questions and attention to survey metadata enable accurate and meaningful analyses.

9.2.1. Primary data producers

14. They are responsible for planning and executing the survey to collect the intended data. For them, planning the analysis typically involves preparing a *tabular plan* — a document specifying the core set of tables to be produced once the survey data becomes available. This plan helps to ensure that the survey results align with the stated needs and objectives of the survey (see Chapter 2). It also helps those designing the survey questionnaire(s) to ensure that all items needed are included, and possibly, help avoid adding items / questions that are not really needed for the intended analysis (see Chapter 4).

15. Preparing a *tabular plan* generally requires defining three sets of specifications:

[1] **Filter Conditions**: These may be used to define subgroups of the population for which specific tables will be produced. For example, in a survey where questions regarding occupation are asked

only from individuals aged 15 or older, a filter condition might be 'if age > 14'. Such a condition means that only those in the relevant age group would be included in tables for the occupation related variables (status, type, income, etc.).

[2] **Classification or Domain Variables**: These are variables used to subdivide the population into meaningful groups for analysis. For example, geographic areas (e.g., states or provinces), age groups, or sex might define rows in a table. These variables are often chosen to meet reporting requirements, such as providing estimates by province in national household surveys. A typical list of domain defining variables would include: geographic levels (provinces, etc.); type of region (urban x rural); sex; age groups; education; race / ethnicity; etc.

[3] **Response or Survey Variables**: These are the variables being analysed to understand how they vary across the defined domains. For instance, continuous survey variables (like income) might be used to create columns in a table, summarizing means, medians, or other statistics. These variables are all others that we do not use as classification ones. Categorical survey variables (like employment status) will generate a column for each category where the corresponding cross-classified frequencies will appear, with one row for each of the classes of the domain defining variables.

16. As discussed throughout Chapter 7, an important consideration when defining domains is related to sample design, particularly when defining strata and sample sizes. For most national household sample surveys, providing breakdowns by province or state is required, and therefore stratification by province will be essential. Additionally, if precision requirements must be met at the provincial level, sample sizes that satisfy these requirements must be computed for each province and then summed to obtain the national sample size.

17. For instance, for domains defined by characteristics that are unavailable from the sampling frame, say age groups, for the case of household surveys that use area sampling, sample size calculations must take into account what the required total sample size must be such that estimates for the rarest group meet precision specifications. As an example, suppose that estimates are required by age groups such as young adults (18 to 29), adults (30 to 49), ageing adults (50 to 59) and elderly (60 and over). Assuming that the population distribution by these age groups is such that the ageing adults is the rarest group with 12.5% of the total population, and if a minimum sample size of 500 individuals in this group is required, then the total sample must be at least 500 / 0.125 = 4,000. That is, in order for the full sample to provide an expected sample of about 500 ageing adults we must sample at least 4,000 individuals for the survey.

9.2.2. Secondary data users

18. Secondary data users are all those who will access and analyse the survey data after it has been released, typically with access only to the public datasets and documentation provided by the data producers or curators. Their first task is to define clear research questions and locate the relevant survey data and documentation (metadata). Good survey metadata must describe the sampling design and estimation methods used (including details about stratification, clustering, and survey weights), both for descriptive parameters and for the corresponding measures of precision.

19. For example, based on the aforementioned, a well-defined research question posed by some secondary data user, can be such as **"Do rural households face digital exclusion compared to urban households?"**. Notice that it directs the analysis towards estimating relevant parameters and precision measures.

20. The question might involve testing whether the proportion of rural households with internet access is significantly lower than that of urban households. This clarity allows for more direct and accurate analyses. The null hypothesis behind this question is that $H_0) P_{Rural} = P_{Urban}$, which we wish to test against the alternative $H_A) P_{Rural} < P_{Urban}$, where P denotes the proportion of households having internet access in the domain identified.

21. A related but different set of hypotheses could be $H_0) C_{Rural} = C_{Urban}$ versus $H_A) C_{Rural} > C_{Urban}$ where C denotes the average cost of connection for broadband internet access in the specified domain. Listing the research questions in this way would enable the subsequent analysis to progress more directly towards the estimation of the relevant parameters from the survey data, and corresponding precision measures, both of which are required to compute test statistics that would provide the evidence required to answer them.

22. In order for such estimation to take place, the secondary data user must first find out details about how the sampling was conducted so that the user can incorporate the sampling design features provided in the database during the analysis stage. As we will discuss in subsequent sections of this chapter, it is essential to account for the survey weights (see Chapter 8) when computing point estimates of both descriptive and model parameters, and to account for the structural components of the sampling design and estimation process (stratification, clustering, unequal inclusion probabilities, non-response adjustment, and calibration of survey weights, if any) when estimating variances or other measures of precision for the point estimates.

23. Users that disregard such aspects of the sampling design do so at their peril, and may end up producing biased estimates that will lead to incorrect inferences and decisions. The recommended practice is for data producers to provide sufficient detail about such sampling aspects as part of the metadata released with survey microdata, in order to enable secondary data users to consider these aspects when conducting their analyses of interest.

9.2.3. Quality control for secondary users

24. A standard quality control process for secondary data users would consist of the following steps:

[1] **Load the data and metadata**: Ensure that the survey microdata is properly linked to metadata describing the key aspects of the sampling design (stratification and clustering identifiers, unequal probabilities of selection) and estimation processes (weights, replicate weights if any, non-response correction, calibration, etc.).

[2] **Replicate published estimates**: Recreate some of the estimates provided by the survey producers, including measures of precision, to confirm that the data and sampling design aspects have been correctly interpreted and loaded.

[3] **Compare scenarios**: When the user is capable of replicating published estimates, she can then proceed towards the required new analysis for which no previous results are available. Repeat this analysis under two scenarios: ignoring and accounting for the sampling design. Compare the results and assess the impact of incorporating the sample design.

[4] **Finalize the analysis**: Use only the results that account for the sampling design in the final interpretation, ensuring that the design effects are appropriately incorporated.

9.3. Accounting for the sampling design

25. When analysing household survey data, ignoring the sampling design undermines the representativeness, accuracy, and credibility of survey-based findings, which can lead to incorrect decisions. This is why accounting for the sampling design is essential when analysing household survey data to ensure valid and unbiased estimates. As seen in the previous chapters, regular household surveys have two major characteristics:

- They use *complex sampling designs* (e.g., stratification, clustering, and unequal probabilities of selection) to represent the population efficiently.
- They define *sampling weights* for each sampling unit (primary, and remaining ones) to properly represent the population.

26. To illustrate this fact, we provide a simple example. Suppose a country has two regions: Region A with 100 people, and Region B with 900 people. Wealthy people live in Region A, with an average income of $10,000, less wealthy people live in Region B, with average income of $2,000. The true population average income is $2,800, because:

$$\theta = \frac{(100 \times 10{,}000) + (900 \times 2{,}000)}{100 + 900} = 2{,}800. \tag{9-1}$$

27. Suppose a survey is conducted where 50 people are sampled from each Region. After data collection, it was found that the sample mean for Region A was $10,000, while the sample mean for Region B was $2,000. If the sampling design is ignored, all units in the sample will receive equal weights, regardless of their corresponding population sizes. This way, the average income is estimated with severe bias as:

$$\hat{\theta} = \frac{(50 \times 10{,}000) + (50 \times 2{,}000)}{100} = 6{,}000. \tag{9-2}$$

despite the fact that both sample averages within regions are perfect estimates of the corresponding population averages.

28. When the sampling design is considered, weights are applied proportional to the population sizes of each region. This way, units in Region A would receive a weight of $\frac{100}{50} = 2$, while units in Region B would receive a weight of $\frac{900}{50} = 18$. In this escnario, the average income is unbiasedly estimated as:

$$\hat{\theta} = \frac{(2 \times 50 \times 10{,}000) + (18 \times 50 \times 2{,}000)}{(2 \times 50) + (18 \times 50)} = 2{,}800. \tag{9-3}$$

Ignoring the sampling design (and corresponding weights) causes that Region A (smaller but wealthier) dominates the estimate, even though its population represents only 10% of the country's population. This creates a bias, making the average population income seem higher than it actually is. By considering the sampling design (the sample was stratified by region, with equal allocation), and using weights, the latest estimate correctly reflects the true contribution of each Region to the global average, thus avoiding bias.

29. Another crucial aspect to take into account is clustering. Most household sample surveys select an area sample, with clusters defined by census enumeration areas or similar groups of households. Then from within each sampled cluster, a sub-sample of households is selected for the survey. While such designs are cost efficient, they imply some loss in precision for the resulting estimates.

30. Consider a toy example where clusters are city blocks, and a sample of 10 households is selected within each cluster to be surveyed. Suppose the target parameter is the proportion of households which have electricity provided from a city grid. Assuming that each cluster either is included or excluded from the grid, a sample of 1,000 households obtained from such a design would be as precise as sample of only 100 households, if they were sampled by simple random sampling. The reason is that after observing a single household in each cluster, the other nine we sample there are not bringing in any new information. Any analysis of the full sample of households would fool the user to believe that the estimates are as precise as having sampled 1,000 households at random from the population, which was not the case, and standard errors would be severely underestimated.

31. In real applications, to enable adequate analysis that accounts for the sampling design, survey datasets must contain identifiers for strata and primary sampling units (PSU), as well as weights for relevant units of analysis (households, individuals, etc.). Alternatively, when such information is not available, the dataset should at least contain replicate weights, and the user should have clear guidance on how to compute both point and variance estimates.

32. This section discusses how survey data from a sample can be used to draw conclusions about an entire population using a design-based approach. This approach assumes that the sample is selected through a well-defined probability sampling process, which ensures that every unit in the population has a known, non-zero chance of being sampled. Sampling weights, which reflect how much each sampled unit represents in the population, are essential tools in this approach. These weights allow users to make estimates that account for the sampling process and produce results that are representative of the population. A well-described survey design facilitates statistical analysis, supports effective data interpretation, and enables meaningful insights into complex phenomena. Not accounting for it may lead to biased estimates and misleading conclusions, as illustrated.

9.3.1. Parameters and estimators

33. Two common goals when analysing survey data are to estimate the value of a characteristic for the whole population, such as total income, and to estimate the average value of that characteristic, such as the average income per person. These are referred to as the *population total* and the *population mean*, respectively. The design-based approach incorporates the probability sampling design in the inferential process.

34. Under probability sampling, every unit in the population has a known chance of being included in the sample. The *sample inclusion probabilities* are then used to calculate *basic sampling weights*. When the design is properly incorporated, estimates for population totals and means are unbiased. This means that, on average, the estimates will equal the true population values if the survey were to be repeated many times under the same conditions.

35. The estimators used to make inferences about the population parameters use the sampling weights to create weighted sums of the survey data, which serve as estimates for the population parameters. If the weights are appropriately applied, the resulting estimates are consistent with the true values in the population.

36. As stated in Chapter 8, in some situations, the basic sampling weights need adjustments to improve the accuracy of survey estimates. Adjustments may account for survey non-response, where weights of responding units are corrected to account for units that were selected for the survey but did not participate. These adjusted weights help to minimize biases in the estimates and make the results more reliable. If

calibration was performed, the weights are modified to ensure that the weighted sums align more closely with known population distributions, such as age or sex distributions.

37. The *population total* $Y = \sum_U y_k$ and *mean* $\overline{Y} = \frac{Y}{N}$ of a survey variable y can be estimated by weighted estimators given by $\hat{Y}_{HT} = \sum_s d_k \ y_k$ and $\overline{y}_H = \frac{\hat{Y}_{HT}}{\hat{N}_{HT}} = \frac{\sum_s d_k \, y_k}{\sum_s d_k}$, respectively. The basic design weights $d_k = 1/\pi_k$ are given by the reciprocals of the sample inclusion probabilities of the sampled units (denoted π_k), for all $k \in s$.

38. When the survey weights are calibrated and/or non-response adjusted, the above expressions may still be used, but with the calibrated or non-response adjusted weights, w_k say, replacing the basic design weights d_k, for all $k \in s$.

39. Here $s = \{k_1, \dots, k_n\} \subset U$ denotes the set of units in a sample selected from the population U using a *probability sampling design* $p(s)$, that ensures strictly positive first order inclusion probabilities $\pi_k = Pr(k \in s), \forall \, k \in U$. These inclusion probabilities are assumed known $\forall \, k \in s$, at least to the data producers.

40. An important part of survey analysis is understanding the level of uncertainty in the estimates. Since we are working with a sample and not the entire population, there will always be some variability in the estimates. This variability is measured using either the *sampling variance*, the *standard error* or the *coefficient of variation* (*cv*) of the estimates. The latter two are functions of the first, and serve to indicate by how much the estimate might differ from the true population value in absolute (*se*) or relative (*cv*) terms. Under the design-based framework and assuming full response, $\hat{Y}_{HT}$ is unbiased for Y and its sampling variance can be estimated unbiasedly by

$$\hat{V}_p(\hat{Y}_{HT}) = \sum_{k \in s} \sum_{j \in s} \left(d_k d_j - d_{kj}\right) y_k y_j \tag{9-4}$$

where $d_{kj} = 1/\pi_{kj}$ and $\pi_{kj} = Pr(k, j \in s), \forall \, k, j \in U$. This result assumes that the sampling design $p(s)$ is such that $\pi_{kj} > 0 \ \forall \, k, j \in U$. Considering the sampling design is crucial for the computation of variance. For example, consider a population of size $N = 6$, and a sample of size $n = 3$. Assume that the following sample values were observed: $(y_1 = 10, y_2 = 14, y_3 = 18)$. If the sampling design is simple random sampling without replacement, the above formula becomes:

$$\hat{V}_{SRS}(\hat{Y}_{HT}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S_{y_s}^2 \tag{9-5}$$

where $S_{y_s}^2$ is the sample variance. After a simple algebraic manipulation, the formula becomes $\hat{V}_p(\hat{Y}_{HT}) = \frac{36}{3} \cdot \left(1 - \frac{3}{6}\right) 16 = 96$. A naive analyst might incorrectly compute the variance using the following formula, which ignores the sampling design: $\frac{N^2}{n} S_{y_s}^2 = 192$.

41. Notice that the estimate for the population total is $\hat{Y}_{HT} = 84$. The proper standard error is $\sqrt{\hat{V}_p(\hat{Y}_{HT})} = \sqrt{96} \approx 9.80$. The 95% confidence interval is $84 \pm 19.25 = [64.75, 103.25]$. Using the naive variance, an incorrect 95% confidence interval is $[56.80, 111.20]$. The naive confidence interval is wider than the correct confidence interval, and it overestimates the uncertainty in the total. This example clearly

demonstrates the importance of considering the sampling design when calculating variances and confidence intervals. Ignoring the sampling design can lead to incorrect inferences.

42. While the formula for variance estimation $\widehat{V_p}$ is general and covers the vast majority of sample designs used in the practice of household sample surveys, it is not used in practice because the second order inclusion probabilities π_{kj} (and corresponding pairwise weights d_{kj}) are generally unknown to secondary survey data users. In fact, even data producers do not compute such pairwise weights, since there are more efficient methods for variance estimation that do not require having such weights.

43. For this reason, simpler and more efficient methods are often used in practice, allowing users to quantify the uncertainty without requiring overly detailed information about the sampling design.

9.3.2. Approaches to variance estimation

44. When working with household surveys, the sample is usually only a small subset of the entire population. Because of this, it is important to provide not only the main estimates of interest, such as totals or averages, but also the level of uncertainty in these estimates.

45. Understanding and estimating uncertainty is a critical part of analysing household survey data; by using proper methods, users can measure the reliability of their estimates. There are several methods to estimate the uncertainty in survey results.

- *Estimating Equations* (David A. Binder 1983b), which comprises a unifying idea of sampling theory, provides a flexible framework for estimating totals, means, ratios, and other parameters as well as their corresponding variances.
- *Taylor Linearization* is an approach that relies on approximating complex non-linear statistics by linear ones, and then estimating the variance / se / cv of the linear approximating quantity.
- The *Ultimate Cluster* method is often used in surveys that collect data through stratified multi-stage sampling, and relies on computing the variance between quantities calculated at the level of the primary sampling units (PSU). It is often combined with *Taylor Linearization* for obtaining estimates of variances of non-linear statistics, such as means, ratios, etc.
- The *Bootstrap* and other replication methods rely on sampling repeatedly from the observed sample, computing estimates from each replica, and then using the variability between the estimates to estimate the variance / se / cv of the main estimate.

46. With the help of modern software, all of these approaches can be implemented efficiently, ensuring accurate and meaningful analysis of survey data.

9.3.2.1. Estimating equations

47. Many population parameters can be written/obtained as solutions of *population estimating equations*. Variance estimation for these sample-based methods follows a consistent framework. Although the details can be technical, the key idea is that the same principles used to estimate totals can be applied to estimate variances. This generality makes the method simple and versatile, allowing it to be well implemented in widely used software like the R survey package and the Stata svy functions. These tools automate much of the process, making it accessible for users to estimate both population parameters and their associated uncertainties.

48. A generic population *estimating equation* is given by $\sum_{i\in U} z_i(\theta) = 0$, where $z_i(\cdot)$ is an *estimating function* evaluated for unit i and θ is a population parameter of interest. These equations provide a general way to define and calculate many population parameters, such as totals, means, and ratios. The concept is straightforward: population parameters can be defined as solutions to specific equations that involve all the units in the population. This approach is flexible and can be adapted to calculate many different types of parameters.

- For the case of the population total, take $z_i(\theta) = y_i - \theta/N$. The corresponding population estimation equation is given by $\sum_{i\in U}(y_i - \theta/N) = 0$. Solving for θ gives the population total $\theta_U = \sum_{i\in U} y_i = Y$.
- For ratios of population totals, taking $z_i(\theta) = y_i - \theta x_i$, the corresponding population estimating equation is given by $\sum_{i\in U}(y_i - \theta x_i) = 0$. Solving for θ gives the *population ratio* $\theta_U = \sum_{i\in U} y_i / \sum_{i\in U} x_i = R$.
- Similarly, for population means, take $z_i(\theta) = y_i - \theta$.

49. The idea of defining population parameters as solutions to population estimating equations allows defining a general method for obtaining corresponding sample estimators. It is a matter of using the *sample estimating equations* $\sum_{k\in s} d_k\ z_k(\theta) = 0$. Under *probability sampling*, full response and with $d_k = 1/\pi_k$, the sample sum in the left-hand side is unbiased towards the population sum in the corresponding population estimating equation. Solving the sample estimating equation yields consistent estimators for the corresponding population parameters.

50. A consistent estimator for the variance of non-linear estimators obtained as solutions of sample estimating equations, derived using *Taylor Linearization*, is given by:

$$\hat{V}_{TL}(\hat{\theta}) = [\hat{J}(\hat{\theta})]^{-1}\hat{V}_p\left[\sum_{k\in s} d_k\ z_k(\hat{\theta})\right][\hat{J}(\hat{\theta})]^{-1} \qquad (9\text{-}6)$$

where $\hat{J}(\hat{\theta}) = \sum_{k\in s} d_k\ [\partial z_k(\theta)/\partial\theta]_{\theta=\hat{\theta}}$.

51. This approach implies that one is able to estimate many population parameters and corresponding variances using essentially well-known methods for estimating totals and their variances. Its simplicity and generality have enabled the development of software such as the R survey package, the Stata svy functions and others.

9.3.2.2. Ultimate cluster

52. The *Ultimate Cluster* is a straightforward and powerful approach for estimating the variance of totals in surveys that use stratified multi-stage cluster sampling designs. This method, proposed by Hansen, Hurwitz, and Madow (1953), simplifies the complex nature of multi-stage designs by focusing only on the variation between the largest groups, known as Primary Sampling Units (PSUs). It assumes that, within each sampling stratum, PSUs were sampled independently with replacement (potentially with unequal probabilities), even if they were actually selected without replacement in the actual sampling process.

53. The method considers only the variation between statistics computed at the level of PSUs. This method allows for a simpler variance estimation, while still providing reliable variance estimates. This idea is simple, but quite powerful, because it allows to accommodate a variety of sampling designs, involving stratification and selection with unequal probabilities (with or without replacement) of both PSUs as well

as lower level sampling units (households and individuals). The requirements for the application of this method are the following:

- Unbiased estimates of totals for the variable(s) of interest are available for each sampled PSU.
- Data are available for at least two sampled PSUs in each stratum (if the sample is stratified in the first stage).
- The survey dataset contains all the information regarding PSUs, strata and weights.

54. Consider a multi-stage sampling design, in which m_h PSUs are selected in stratum h, with $h = 1, \dots, H$. Let $\hat{Y}_{hi} = \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} d_{hik}\ y_{hik}$ denote an estimate of the population total Y_h in stratum h based on the single PSU i sampled in this stratum. Then an unbiased estimator of the population total $Y = \sum_{h=1}^{H} \sum_{i \in U_{1h}} Y_{hi}$ is given by $\hat{Y}_{UC} = \sum_{h=1}^{H} \hat{Y}_h$ where $\hat{Y}_h = \frac{1}{m_h} \sum_{i \in s_{1h}} \hat{Y}_{hi}$. The *Ultimate Cluster* estimator of the corresponding variance is given by:

$$\hat{V}_{UC}(\hat{Y}) = \sum_{h=1}^{H} \frac{m_h}{m_h - 1} \sum_{i \in s_{1h}} \left(\hat{Y}_{hi} - \hat{Y}_h\right)^2 \qquad (9\text{-}7)$$

where U_{1h} and s_{1h} are the population and sample sets of PSUs in stratum h, for $h = 1, \dots, H$. (See for example, (Babubhai. V. Shah et al. 1993), p. 4).

55. Although the method was originally proposed for estimation of variances of estimated totals, it can also be applied in combination with *Taylor Linearization* and *Estimating Equations* approaches to obtain variance estimates for estimators of many other population quantities that can be obtained as solutions to sample estimating equations. This makes the method versatile and useful for a wide range of applications in survey analysis.

56. One key assumption of the method is that, within strata, the PSUs were selected independently and with replacement. In reality, most surveys select PSUs without replacement, which provides for more efficient designs. However, the variance estimates produced by the *Ultimate Cluster* method are generally close enough to be useful, even under these conditions, provided the sampling fraction of PSUs is small.

57. The *Ultimate Cluster* method is particularly attractive because of its simplicity. Survey practitioners often prefer it over more complex approaches that account for all stages of the sampling design. Although these detailed methods may provide slightly more accurate variance estimates, they are harder to implement and require more detailed information about the sampling process. In contrast, this method offers a reasonable approximation that works well for most practical purposes, especially for estimating totals or averages. A discussion about Quality of this approximation and alternatives can be found in (Särndal, Swensson, and Wretman 1992), p. 153.

9.3.2.3. Bootstrap

58. Replication methods for variance estimation are based on the idea of re-sampling from the available sample, computing the estimates from each replica, and then using the variability between the estimates across replicas to estimate the variance. They are particularly useful when the user does not have access to information on stratum and/or PSUs identifiers in the database, and the *Ultimate Cluster* method cannot be used in such cases.

59. The *bootstrap* method comprises a powerful and flexible approach for estimating variances in surveys and many other contexts. Originally proposed by Efron (1979), the version commonly used for household surveys is called the Rao-Wu-Yue Rescaling Bootstrap (Rao, Wu, and Yue 1992). This method

is well-suited for stratified multi-stage sampling designs and has become widely used for variance estimation with complex survey data.

60. Conceptually, the *bootstrap* method relies on creating many new "replicated" datasets, which are slightly different versions of the original sample. These replicated datasets mimic the process of repeatedly drawing samples from the population. By analysing the variation in results across these datasets, we can estimate how much uncertainty there is in our estimates from the original sample. In practice, the method can be applied by creating multiple columns of weights in the original sample data set, with weights modified to mimic the process of re-sampling from the available sample. The creation of these multiple columns of weights should be done by the National Statistical Office following the steps outlined below.

[1] First, create a new sample for each stratum by randomly selecting primary sampling units (PSUs) from the original sample, allowing PSUs to be selected more than once (with replacement). Each selected PSU is included in the new dataset along with all its associated data and lower level units. The size of this random sample with replacement is of $m_h - 1$ PSUs in each of the H design strata.

[2] This process of creating new samples is repeated many times, usually hundreds or thousands, to produce multiple "replicated" datasets. That is, repeat Step 1 R times, and denote by $m_{hi}(r)$ the number of times that the PSU i of stratum h was selected for the sample in replicate r.

[3] For each replicate r, *bootstrap* weights are calculated for each unit. These weights account for how often each PSU appears in the replicate and ensure that the replicated datasets remain representative for the population. The *bootstrap* weight for unit k within PSU i of stratum h in replica r is $w_{hik}(r) = w_{hik} \times \frac{m_h}{m_h - 1} \times m_{hi}(r)$.

[4] The parameter of interest, such as a total or mean, is estimated for each replica using the *bootstrap* weights for that replica. That is, for each replica r, calculate an estimate $\hat{\theta}_{(r)}$ of the target parameter $theta$ using the *bootstrap* weights $w_{hik}(r)$ instead of the original weights w_{hik}.

[5] Finally, the variability of the results across all replicas is used to estimate the variance. The idea is that the variation in these replicate estimates reflects the uncertainty in the original estimate. This estimate of the variance takes the following form:

$$\hat{V}_B(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{(r)} - \tilde{\theta}\right)^2 \qquad (9\text{-}8)$$

where $\tilde{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{(r)}$ is the average of the replica estimates.

61. Whenever the original sampling weights w_{hik} receive non-response adjustments or are calibrated, the corresponding non-response adjustments and/or calibration of the basic weights must be repeated for each replica, so that the variance estimates adequately reflect the effects of the calibration and non-response adjustments on the uncertainty of the point estimates. This ensures that the variance estimates accurately reflect the additional uncertainty introduced by these adjustments.

62. The bootstrap method has several advantages. It works well for complex survey designs and can handle a wide range of parameters, including those that are difficult to estimate using traditional methods, such as medians or other nonlinear statistics. It also provides a way to estimate variances when other methods are not available or practical to use. The method is particularly helpful for users analysing a survey

that does not provide the corresponding design variables (strata and PSUs) but does provide a set of replicated weights. Notice that, given the simplicity of the method, users should not feel restricted to using specialized software for calculating variances under this approach.

63. Many modern statistical software tools, including the survey package in R, support bootstrap replication and variance estimation, making it accessible to a wide range of users. While the bootstrap method is computationally intensive, requiring many replicas to be created and processed, it is highly effective. It provides robust variance estimates even for complex parameters and remains one of the most flexible tools for analysing survey data.

9.3.3. Using software to generate valid inferences

64. The design and analysis of household surveys must make extensive use of existing computational tools. This section reviews some computational approaches within statistical software that are used for each of the statistical processes required to publish official figures with appropriate levels of accuracy and reliability. Key processes that analysts should focus on include: modelling nonresponse and statistical imputation, estimating standard errors for each indicator of interest to be included in the production tables, and analysing multivariate relationships between survey variables.

65. Nations (2005, sec. 7.8) highlights the importance of including the structure of complex survey designs in the inference process for estimating official statistics from household surveys. It warns, with an empirical example, that failing to do so may result in biased estimates and underestimated sampling errors. Below are some key features that statistical software packages incorporate when managing data from complex survey designs, such as those found in household surveys. A more detailed review, including syntax and computational code, can be found in Steven G. Heeringa, West, and Berglund (2017a, Appendix A).

66. In general, these computational tools are designed to enhance the efficiency of variance approximation methods for complex samples, as well as replication techniques to estimate design-based variances (Westat 2007). Some of these software packages are free to use, although most are licensed products requiring paid licenses. These products, in addition to providing descriptive statistics (such as means, totals, proportions, percentiles, and ratios), allow for fitting linear and logistic regression models. All resulting statistics are based on the survey design.

67. Software packages designed for survey analysis often report the design effect ($DEFF$) when processing survey data. This estimate provides a critical measure of how the complexities of the sampling design—such as clustering, stratification, and unequal weights—affect the precision of estimates compared to a simple random sample (SRS) of the same size. By including DEFF in their output, these tools enable researchers to assess the efficiency of their sampling designs and interpret the variability in their data more accurately.

68. As defined by Kish (1965, 258), the $DEFF$ is the ratio between the actual variance of a complex sample and the variance of a simple random sample (SRS) of the same size. This measure is estimated as:

$$DEFF = \frac{\hat{V}_p(\hat{\theta})}{\hat{V}_{SRS}(\hat{\theta})} \tag{9-9}$$

where $\hat{V}_p(\hat{\theta})$ represents the estimated variance of an estimator $\hat{\theta}$ under a complex sampling design $p(s)$, and $\hat{V}_{SRS}(\hat{\theta})$ denotes the estimated variance of the same estimator under a simple random sampling design. The design effect measures the clustering effect introduced by using a complex sampling design compared

to SRS for inferring a finite population parameter θ. According to Nations (2008, 49), the design effect can be interpreted in three ways: as the factor by which variance increases under a complex design compared to SRS, as an indicator of how much less efficient the complex design is in terms of precision, or as a reflection of how much larger the sample size would need to be under the complex design to match the variance achieved by SRS. Park et al. (2003) proposes that the design effect of a survey can be decomposed into three multiplicative components:

[1] The effect due to unequal weighting: this component tends to slightly increase variance if sampling weights are unequal. Uniform weights avoid such increases, making self-weighting designs desirable for household surveys.

[2] The effect due to stratification: this component reduces variance when stratification is optimal, though the reduction is usually modest.

[3] The effect due to multi-stage sampling: this component typically increases the variance of survey estimates because units within the same cluster tend to be more similar to each other than to units in other clusters.

69. In practice, $DEFF$ is especially useful when evaluating the quality of survey estimates and planning future surveys. For instance, a large $DEFF$ indicates that the complex design introduces significant clustering effects or inefficiencies, which can inflate variances and reduce the precision of key estimates. Conversely, a $DEFF$ close to unity suggests that the design features have minimal impact on variance. Understanding these effects allows researchers to decide whether adjustments to weighting, stratification, or sampling stages are needed in future data collection efforts.

70. Statistical software packages such as Stata, R, SAS, and SPSS automatically calculate the design effect within their survey analysis modules. These computations require users to input specific details about the survey design, including sampling weights, strata, and cluster identifiers. Next, we provide a non-comprehensive summary of features and capabilities available in major statistical software.

9.3.3.1. R

71. R is a free software increasingly used in social research, as it is likely to host the latest scientific findings implemented in this software (R Core Team 2024). Being open-source, researchers can upload their own collections of computational functions to the official repository (CRAN) and make them available to the community. The samplesize4surveys package (Rojas 2020) determines the sample size for individuals and households in repeated, panel, and rotational household surveys. The sampling (Tillé and Matei 2016) and TeachingSampling (Gutiérrez 2015) packages enable the selection of probabilistic samples from sampling frames under a wide variety of designs and algorithms. The survey package (Lumley 2016), once the survey design is predefined using the svydesign() function, allows for analysing household survey data and obtaining appropriate standard error estimates.

9.3.3.2. STATA

72. The svy environment provides tools for appropriate inference from household surveys (STATA 2017). The svyset command specifies variables identifying survey design features, such as sampling weights, clusters, and strata. The svydescribe command produces tables describing strata and sampling units at a given survey stage. Once survey design definitions are loaded, any model can be estimated, and the resulting statistics will be survey-design-based. The svy environment also supports predictive commands.

9.3.3.3. SPSS

73. The complex samples module in SPSS (IBM 2017) supports the selection of complex samples through user-defined sampling schemes. Next, an analysis plan must be created by assigning design variables, estimation methods, and sample unit sizes. Once the sampling plan is defined, the module enables the estimation of counts, descriptive statistics, and crosstabulations. It is also possible to estimate ratios and regression coefficients in linear models, along with corresponding hypothesis test statistics. Finally, the module allows for estimating nonlinear models, such as logistic regressions, ordinal regressions, or Cox regressions.

9.3.3.4. SAS

74. This statistical software includes a procedure for selecting probabilistic samples called SURVEYSELECT, which integrates common selection methods such as simple random sampling, systematic sampling, probability proportional to size sampling, and stratified allocation tools. To analyse data from complex samples, specific procedures have been programmed (SAS 2010):

- SURVEYMEANS: Estimates totals, means, proportions, and percentiles, along with their respective standard errors, confidence intervals, and hypothesis tests;
- SURVEYFREQ: Estimates descriptive statistics (e.g., totals and proportions) in one- and two-way tables, provides sampling error estimates, and analyses goodness-of-fit, independence, risks, and odds ratios;
- SURVEYREG and SURVEYLOGISTIC: Fit linear and logistic regression models, respectively, estimating regression coefficients with associated errors and providing an exhaustive analysis of model properties;
- SURVEYPHREG: Fits survival models using pseudo-maximum likelihood techniques.

9.4. Descriptive parameters

75. Descriptive parameters are the most commonly analysed outputs from household survey data. These analyses focus on summarizing key characteristics of the population by estimating values for a variety of survey variables. The goal is to provide clear and meaningful insights into the population using data collected from a representative sample.

76. The most basic and frequently estimated parameters include **frequencies**, **proportions**, **means**, and **totals**. Means and totals provide average and cumulative values, respectively, which are useful for understanding population-level behaviours and trends. Frequencies can show the number of households/people in a specific category (e.g. number of poor people), while proportions can represent the share of households/people meeting a particular condition (e.g. poverty rate).

77. In recent years, the scope of descriptive analysis has expanded beyond these basic parameters. Analysts now estimate more complex metrics, such as **quantiles** of numeric variables, which help describe the distribution of values (e.g., the median income of households). There are also metrics for particular types of analysis, such as poverty (FGT indices), inequality (Gini, Theil, Atkinson), polarization (Wolfson, DER), etc. - see Jacob, Damico, and Pessoa (2024).

9.4.1. Frequencies and proportions

78. One of the most fundamental tasks in household survey analysis is estimating the size of subpopulations, namely the number of people or households in specific categories, as well as the

proportions they represent within the population. These estimates are crucial because they provide a snapshot of the demographic and socioeconomic profile of a population. Policymakers and planners use this information to make decisions about resource allocation, public policy design, and the development of social programs.

79. For example, understanding how many people live below the poverty line, how many are unemployed, or how many have completed a certain level of education provides valuable insights. These insights help address inequalities, support the design of targeted interventions, and promote equitable development across communities. The ability to understand the distribution across categories provides valuable information to address inequalities and promote equitable development.

80. To estimate the size of a population or subpopulation, analysts focus on categorical variables, which divide the population into distinct groups. For example, categories could represent different income quintiles, employment statuses, or education levels. The size of a population refers to the total number of individuals or households in the survey data who fall into a specific category. Population size estimates are calculated by combining the information collected from survey samples with *sampling weights*. These weights indicate how many people or households each surveyed unit represents in the broader population. A sampling estimator of a population size is given by the following expression:

$$\widehat{N} = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \tag{9-10}$$

where s_{hi} is the sample of households or individuals in PSU i of stratum h; s_{1h} is the sample of PSUs within stratum h; and w_{hik} is the weight (expansion factor) of unit k within PSU i in stratum h.

81. Subpopulation size estimates work similarly but focus on a subset of the population defined by a specific characteristic. For example, if we want to estimate the number of people in a particular category, we would identify the relevant group in the survey data and sum up their weights. This approach allows analysts to estimate not only the total population size but also the size of specific groups of interest. This way, a binary variable should be defined, $I(y_{hik} = d)$. It will take the value one if unit k from PSU i in stratum h belongs to category d in the discrete variable y. A sampling estimator for this parameter is given by the following expression:

$$\widehat{N}_d = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\ I(y_{hik} = d) \tag{9-11}$$

82. Proportions describe the relative size of specific groups within the population. For instance, the proportion of households living below the poverty line is a critical measure for understanding socioeconomic disparities. To estimate a proportion, analysts calculate the weighted average of the binary variable. This approach ensures that the estimate accurately reflects the population distribution. As mentioned by Steven G. Heeringa, West, and Berglund (2017b), by recoding the original response categories into simple indicator variables y with possible values of 1 and 0 (e.g., 1=Yes, 0=No), the estimator for a proportion is defined as follows:

$$\hat{p}_d = \frac{\widehat{N}_d}{\widehat{N}} = \frac{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\ I(y_{hik} = d)}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}} \tag{9-12}$$

As this defines a nonlinear estimator, we can apply Taylor linearization to obtain the approximate variance of the above estimator by defining the corresponding estimating function as $z_{hik} = I(y_{hik} = d) - \hat{p}_d$. Many statistical packages provide proportion estimates and standard errors on a percentage scale.

83. When the target proportions are close to 0 or 1, special methods are used to ensure confidence intervals remain meaningful; notice that the limits of the traditional symmetric normal confidence intervals may fall outside the permissible range for proportions. This would have no interpretation due to the nature of the parameter. To address this issue, alternative confidence interval estimates, as proposed by Rust, Hsu, and Westat (2007) and Dean and Pagano (2015) are available. One alternative based on using the logit transformation of the estimated proportion is:

$$CI(\hat{p}_d\ ;\ 1-\alpha) = \frac{exp\left[ln\left(\frac{\hat{p}_d}{1-\hat{p}_d}\right) \pm \frac{t_{1-\alpha/2,\,df} \times se(\hat{p}_d)}{\hat{p}_d(1-\hat{p}_d)}\right]}{1 + exp\left[ln\left(\frac{\hat{p}_d}{1-\hat{p}_d}\right) \pm \frac{t_{1-\alpha/2,\,df} \times se(\hat{p}_d)}{\hat{p}_d(1-\hat{p}_d)}\right]} \tag{9-13}$$

9.4.2. Totals, means and ratios

84. In household surveys, analysing numerical data often involves estimating key descriptive measures such as means, totals, and ratios. These measures summarize important characteristics of the population and provide valuable insights for decision-making. The estimation process can be applied to the entire population or specific subgroups, depending on the research objectives. As mentioned by Steven G. Heeringa, West, and Berglund (2017a), the estimation of population totals or averages for a variable of interest, along with the estimation of corresponding variances, has played a crucial role in the development of probability sampling theory.

85. The estimation of population totals is a fundamental task in survey analysis. Note that population means, proportions and ratios are all dependent on population totals. A total represents the sum of a specific variable (e.g., total income or total expenditure) across the entire population. For example, if the goal is to estimate the total income of all households in a country, we combine data from the sample using weights that account for the survey design and ensure representativeness. For single numeric survey variables, the simplest estimates are for totals and means. Ratios are often used to obtain summaries that relate two numeric variables. Estimates for such parameters can be obtained either for the entire population or disaggregated by domains of interest, depending on the research needs.

86. Once the sampling design is defined, which was done in the previous section, the estimation process for the parameters of interest is carried out. For the estimation of totals with complex sampling designs that include stratification ($h = 1,2,\ldots,H$) and subsampling in PSUs (assumed to be within stratum h) indexed by $i = 1,2,\ldots,m_h$, the estimator for the population total can be written as:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\ y_{hik} \tag{9-14}$$

Under full response, the Ultimate Cluster variance estimator for $\hat{Y}$ was provided in Section 9.2. Modern statistical tools, such as the survey package in R, make it straightforward to calculate totals and their associated uncertainties.

87. The confidence interval of level $1-\alpha$ for the population total Y is given by:

$$\hat{Y} \pm z_{1-\alpha/2} \times \sqrt{\hat{V}_{UC}(\hat{Y})} \tag{9-15}$$

with $z_{1-\alpha/2}$ denoting the quantile of the Gaussian distribution leaving an area of $\alpha/2$ to its right.

88. Population means, or averages, are also very important and provide an understanding of the central tendency of a variable. For instance, the average income of households can indicate the general economic well-being of a population. A mean is calculated as the total of a variable divided by the population size. Since estimating a mean involves both totals and population sizes, the accuracy of a mean estimate depends on the accurate estimation of both components. Specialized techniques, such as resampling methods or Taylor linearization, are used to estimate the uncertainty associated with means. The estimation of the population means is a very important task in household surveys. An estimator of the population mean can be written as the ratio of two estimated population totals, as follows:

$$\widehat{\bar{Y}} = \frac{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\ y_{hik}}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}} = \frac{\hat{Y}}{\widehat{N}}. \tag{9-16}$$

Since $\widehat{\bar{Y}}$ is a nonlinear estimator, there is no closed-form formula for exact the variance of this estimator. For this reason, either resampling methods or Taylor series approximations must be used. The latter may be achieved remembering that for the survey mean the sampling estimating equation requires defining $\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\,(y_{hik} - \theta) = 0$, therefore we can apply the variance estimator given in Section 9.2 with $z_{hik} = y_{hik} - \widehat{\bar{Y}}$.

89. Ratios provide insights into the relationship between two variables. For example, the ratio of household expenditures to income can reveal patterns in spending behavior. A ratio is calculated by dividing one total by another, such as total expenditures by total income. The accuracy of a ratio depends on the precise estimation of both totals. Ratios are particularly useful for creating indicators that help compare groups or track progress over time. As another example, SDG indicator N.2.1.1 is defined as the prevalence of undernourishment. This indicator can be estimated using a ratio of two continuous variables: food consumption (measured in calories or energy intake) and dietary energy requirements (calculated based on factors like age, sex, and physical activity level).

90. Since a ratio is the quotient of two estimators of totals, both the numerator and the denominator are unknown quantities and thus need to be estimated. The point estimator for a ratio in complex surveys is the quotient of the estimators for the totals, as defined by:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\ y_{hik}}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}\ x_{hik}} \tag{9-17}$$

For variance estimation, all you need to do is specify the estimating function as $z_{hik} = y_{hik} - \hat{R}\ x_{hik}$, when y and x are the numerator and denominator variables, respectively, and apply the variace estimator given in Section 9.2.

9.4.3. Correlations

91. Correlation analysis is a useful method for understanding the relationship between two numeric variables in survey data. For example, you might be interested in knowing whether household income and expenditure are related, and if so, how strongly. The Pearson correlation coefficient is commonly used to measure this relationship as it quantifies the strength and direction of a linear relationship between two variables. Its value ranges from –1 to 1:

- A **positive value** indicates that as one variable increases, the other also tends to increase.
- A **negative value** indicates that as one variable increases, the other tends to decrease.
- A value close to **zero** suggests little to no linear relationship between the variables.

92. When analysing survey data, the correlation is estimated using the survey weights. These weights ensure that the estimated correlation reflects the relationships in the entire population, not just the sample. Weighted correlations adjust for the complex survey design, accounting for stratification, clustering, and unequal probabilities of selection. To compute the correlation coefficient, we look at how the variables vary together (their covariance) and normalize this by their individual variations. This normalized measure ensures the correlation is unaffected by the units of measurement of the variables, making it easier to interpret.

93. The Pearson correlation coefficient between two numeric survey variables, say x and y, can be estimated using

$$\hat{\rho}_{xy} = \frac{\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}\left(y_{hik}-\hat{\bar{Y}}\right)\left(x_{hik}-\hat{\bar{X}}\right)}{\sqrt{\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}\left(y_{hik}-\hat{\bar{Y}}\right)^2}\sqrt{\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}\left(x_{hik}-\hat{\bar{X}}\right)^2}} \quad (9\text{-}18)$$

Modern statistical software, such as R, provides functions to calculate weighted Pearson correlation coefficients directly. Tools like the survey package ensure that the correlations are estimated correctly, accounting for the survey design. This allows analysts to obtain accurate and meaningful measures of association.

9.4.4. Percentiles and inequality measures

94. Percentiles and quantiles are useful tools for analysing the distribution of data beyond just the average. These measures divide data into segments to show how values are spread. For example, the 10th percentile indicates the value below which 10% of the data falls, while the median (50th percentile) divides the data into two equal halves. These measures help describe not only central tendencies but also the spread and variation within a dataset. For instance, identifying the top 10% of income earners might guide tax policy, while finding the bottom 15% could inform subsidy programs. The estimation of percentiles relies on the cumulative distribution function (CDF), which represents the proportion of the population with values less than or equal to a given number. Once the CDF is calculated using survey data and weights, percentiles and quantiles can be derived. The CDF for a variable y in a finite population of size N is defined as follows:

$$F(t) = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in U_{1h}}\sum_{k\in U_{hi}} I\left(y_{hik}\le t\right) \quad (9\text{-}19)$$

where $I(y_{hik} \le x)$ is an indicator variable taking the value 1 if y_{hik} is less than or equal to a specific value t, and 0 otherwise. An estimator of the CDF in a complex sampling design is given by:

$$\hat{F}(t) = \frac{\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}\, I(y_{hik}\le t)}{\sum_{h=1}^{H}\sum_{i\in s_{1h}}\sum_{k\in s_{hi}} w_{hik}} \quad (9\text{-}20)$$

Once the CDF is estimated using the survey design weights, the q-th quantile of a variable y is the smallest value of y such that the CDF is greater than or equal to q. As is well known, the median is the value where

the CDF is greater than or equal to $1/2$. Thus, the estimated median is the smallest value where the estimated CDF is greater than or equal to $1/2$. Following Steven G. Heeringa, West, and Berglund (2017b), to estimate quantiles, one first considers the order statistics denoted as $y_{(1)}, \ldots, y_{(n)}$ and finds the value of j $(j = 1, \ldots, n)$ such that:

$$\hat{F}(y_{(j)}) \leq q \leq \hat{F}(y_{(j+1)}) \tag{9-21}$$

Hence, the estimator of the q-th quantile $y_{(q)}$ is given by:

$$\hat{y}_{(q)} = y_{(j)} + \frac{q - \hat{F}(y_{(j)})}{\hat{F}(y_{(j+1)}) - \hat{F}(y_{(j)})}(y_{(j+1)} - y_{(j)}) \tag{9-22}$$

Quantiles are inherently nonlinear measures, making their variance estimation more complex. Kovar, Rao, and Wu (1988) present results from a simulation study where they recommend using the *Balanced Repeated Replication* (BRR) technique.

95. Economic inequality is a critical area of focus for governments and international organizations. The **Gini coefficient** is a widely used measure to quantify inequality in income or wealth distributions. It is derived by comparing the income distribution of a target population to a perfectly equal distribution. In household surveys, it is calculated using weights that account for the survey design, ensuring representativeness. A normalized version of these weights is often used to simplify the calculations. The Gini coefficient ranges from 0 to 1, where 0 indicates perfect equality (everyone has the same income) and values closer to 1 indicate greater inequality. The Gini coefficient is critical for tracking changes in income distribution over time and comparing inequality levels across regions or countries.

96. Following the estimating equation proposed by David A. Binder and Kovacevic (1995), the estimator for the Gini coefficient is given by:

$$\hat{G} = \frac{2 \times \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w^*_{hik} \hat{F}_{hik}\, y_{hik} - 1}{\hat{\bar{Y}}} \tag{9-23}$$

where w^*_{hik} is a normalized sampling weight, defined as

$$w^*_{hik} = \frac{w_{hik}}{\sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik}} \tag{9-24}$$

and $\hat{F}_{hik}$ represents the estimated CDF for individual k in cluster i of stratum h. Osier (2009) and Langel and Tillé (2013) provide important computational details for estimating the variance of this complex estimator.

9.4.5. NSO – Practical example

97. [[In this subsection a NSO will share how they do disseminate its results on basic descriptive statistics, how they publish the resulting tables and how do they deal with the suppression of estimates that do not reach expected quality.]]

9.5. Associations between categorical variables

9.5.1. Motivation and concepts

98. Household surveys often collect data on categorical variables, such as employment status, educational attainment, or access to services. Understanding whether two categorical variables are related,

or *associated*, is an important aspect of survey analysis. For example, are employment status and access to the internet connected in a meaningful way? Categorical variables are those that divide the population into distinct groups or categories. For example:

- **Employment status** might have categories like "employed," "unemployed," and "not in the labour force."
- **Educational attainment** might include categories such as "primary," "secondary," and "tertiary."

99. This section introduces methods to describe and infer associations between pairs of categorical variables. When analysing associations between two categorical variables, we are interested in whether the distribution of one variable depends on the categories of the other. To assess the relationship between two categorical variables, analysts examine how often different combinations of categories occur. For example, they might count how many individuals fall into each pairing of employment status and educational attainment. These counts are then used to calculate proportions, which describe the relative frequency of each pairing within the population.

100. Analysing associations between categorical variables is useful in various contexts, such as policy development, where understanding the relationship between education and employment helps design effective workforce policies; program evaluation, where assessing whether access to healthcare varies by income level can inform targeted interventions; and social research, where studying connections between demographic factors and access to services provides insights into societal trends.

101. In practice, this analysis often starts with a **contingency table**, a grid that shows the counts or proportions of units in each combination of categories for the two variables. For example, one axis of the table might list employment statuses, while the other lists levels of educational attainment.

102. We start by defining some notation. Let x and y denote two categorical variables, having R and C classes respectively. In order to formulate hypothesis tests for the independence between x and y, we need to consider a *superpopulation model*. We assume that the pairs (x_{hik}, y_{hik}) correspond to observations from identically distributed random vectors $(X; Y)$, that have joint distribution specified by

$$P_{rc} = Pr(X = r\ ;\ Y = c) \quad \text{for } r = 1, \ldots, R \text{ and } c = 1, \ldots, C \tag{9-25}$$

with $\sum_r \sum_c P_{rc} = 1$.

103. If a census could be carried out to collect data on x and y from every unit in the population, we could calculate the population counts of units having classes (r, c) for (x, y) given by:

$$N_{rc} = \sum_{h=1}^{H} \sum_{i \in U_{1h}} \sum_{k \in U_{hi}} I\ (x_{hik} = r\ ;\ y_{hik} = c) \tag{9-26}$$

and the corresponding population proportions as $p_{rc} = N_{rc}/N_{++}$, where $N_{++} = \sum_r \sum_c N_{rc}$ denotes the total number of units in the population.

104. Under the superpopulation model, the population proportions p_{rc} could be used to estimate (or approximate) the unknown probabilities P_{rc}. Since in most instances we will have samples, not censuses, the population proportions p_{rc} must be estimated using weighted estimators provided in the previous sections.

9.5.2. Cross-tabulations

105. Cross-tabulations, also known as contingency tables, are a fundamental tool in survey analysis. They organize data into a table format, showing the frequency distribution of two or more categorical variables. By summarizing relationships between these variables, cross-tabulations help researchers identify patterns and associations that might otherwise go unnoticed. This type of analysis is widely used in research and policy decision-making, as it provides a straightforward way to explore how different variables interact. For example, a contingency table might examine how employment status varies by educational attainment, or how access to the internet differs between urban and rural households.

106. Procedures for assessing independence can be used to determine whether the cross-classified variables are related or independent. This type of analysis is important in many research and decision-making settings. In the specialized literature, cross-tabulations are also referred to as *contingency tables*. Here a table is a two-dimensional array with rows indexed by $r = 1, \dots, R$ and columns indexed by $c = 1, \dots, C$. Such tables are widely used in household survey analysis as they summarize the relationship between categorical variables in terms of frequency counts.

107. A contingency table aims to succinctly represent the association between different categorical variables. It is a grid with rows and columns that represent the categories of two variables. Each cell in the table contains the frequency or proportion of observations that fall into the corresponding combination of categories. The rows might represent categories of a domain defining variable such as "education level" (primary, secondary, tertiary). The columns might represent categories of another variable, such as "employment status" (employed, unemployed, not in the labour force). The table can also include **marginal totals**, which summarize the data for each row or column, and a **grand total**, representing the overall population.

108. In household surveys, frequencies in contingency tables are calculated using survey weights. These weights ensure that the estimates accurately reflect the entire population, accounting for the sampling design. For each cell, the weighted frequency represents the estimated number of individuals in the population with the corresponding combination of categories. For instance: we consider the case of a two-way contingency table. For most household sample surveys, a typical tabular output comprises the weighted frequencies that estimate the population frequencies, as shown in **Figure 9.1**:

Figure 9.1. Contingency table with the weighted frequencies that estimate the population frequencies.

	y			
x	1	...	C	row marg.
1	$\widehat{N}_{11}$	...	$\widehat{N}_{1C}$	$\widehat{N}_{1+}$
...	...	$\widehat{N}_{rc}$	...	...
R	$\widehat{N}_{R1}$	...	$\widehat{N}_{RC}$	$\widehat{N}_{R+}$
col. marg.	$\widehat{N}_{+1}$	...	$\widehat{N}_{+C}$	$\widehat{N}$

where the estimated frequency in cell (r, c) is obtained as

$$\widehat{N}_{rc} = \sum_{h=1}^{H} \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \; I(x_{hik} = r \; ; \; y_{hik} = c) \tag{9-27}$$

and $\widehat{N}_{r+} = \sum_c \widehat{N}_{rc}$, $\widehat{N}_{+c} = \sum_r \widehat{N}_{rc}$ and $\widehat{N}_{++} = \sum_r \sum_c \widehat{N}_{rc}$.

109. Weighted frequencies can also be converted into **proportions**, which indicate the relative size of each group compared to the total population. Proportions are particularly useful when comparing groups of different sizes or when focusing on the relative distribution of categories. The estimated proportions from these weighted sample frequencies are obtained as follows:

$$\hat{p}_{rc} = \frac{\widehat{N}_{rc}}{\widehat{N}_{++}} \tag{9-28}$$

and $\hat{p}_{r+} = \widehat{N}_{r+}/\widehat{N}_{++}$ and $\hat{p}_{+c} = \widehat{N}_{+c}/\widehat{N}_{++}$.

110. Two-way tables can also display the estimates of population relative frequencies (**Figure 9.2**).

Figure 9.2. Contingency table with estimates of population relative frequencies.

	y			
x	1	...	C	row marg.
1	$\hat{p}_{11}$	...	$\hat{p}_{1C}$	$\hat{p}_{1+}$
...	...	$\hat{p}_{rc}$	...	...
R	$\hat{p}_{R1}$	...	$\hat{p}_{RC}$	$\hat{p}_{R+}$
col. marg.	$\hat{p}_{+1}$	...	$\hat{p}_{+C}$	1

111. While tables are a clear way to present data, visualizations such as stacked bar charts or heatmaps can enhance understanding by highlighting patterns and differences between categories. These visuals complement contingency tables, making it easier to communicate findings to a broad audience. More on this will be elaborated in Section 9.7.

9.5.3. Testing for independence

112. In household surveys, it is often important to determine whether two categorical variables are associated or independent (i.e., whether the distribution of one variable is unaffected by the categories of the other). For example, is there a relationship between "educational level" and "employment status"? To answer such questions, *independence tests* are used. These tests compare the observed data with what would be expected if the two variables were completely unrelated.

113. To perform these tests, analysts rely on models that assume the data comes from a larger, hypothetical population (a *superpopulation*). The observed data from the survey is treated as a sample from this superpopulation, and the analysis aims to draw conclusions about the larger population. The starting point for testing independence is the **null hypothesis**, which assumes that the two variables are independent. This

means the likelihood of being in any combination of categories is simply the product of their marginal probabilities.

114. To test this hypothesis, observed frequencies (or proportions) in a contingency table are compared with the expected frequencies under the null hypothesis. If the observed and expected values differ significantly, the null hypothesis of independence is rejected, suggesting an association between the variables. This way, it is possible to perform independence tests to verify whether x and y are associated. Following Steven G. Heeringa, West, and Berglund (2017b), the null hypothesis that x and y are independent is defined as:

$$H_0) \;\; P_{rc}^0 = P_{r+} \times P_{+c} \;\; \forall\, r = 1, \dots, R \text{ and } c = 1, \dots, C. \tag{9-29}$$

115. Hence, to test the independence hypothesis we compare the estimated proportions $\hat{p}_{rc}$ with the estimated expected population proportions under the null P_{rc}^0. If there is a large difference between them, then the independence hypothesis would not be supported by the data.

116. Testing for independence in survey data requires adjustments to account for the sampling design, which often includes stratification, clustering, and unequal probabilities of selection. The **Rao-Scott adjustment** modifies traditional chi-square tests to incorporate these design effects. The test statistic is adjusted for the survey design using a measure called the **generalized design effect (GDEFF)**, which accounts for the complexity of the sampling design. It follows a chi-square distribution under the null hypothesis. Therefore, the following Pearson Rao-Scott adjusted test statistic X_{RS}^2 (Rao and Scott 1984) is defined by:

$$X_{RS}^2 = \frac{n_{++}}{GDEFF} \sum_r \sum_c \frac{\left(\hat{p}_{rc} - \hat{P}_{rc}^0\right)^2}{\hat{P}_{rc}^0} \tag{9-30}$$

where $\hat{P}_{rc}^0 = \hat{p}_{r+} \times \hat{p}_{+c}$ estimates the cell frequencies under the null hypothesis and $GDEFF$ is an estimate of the generalized design effect (Steven G. Heeringa, West, and Berglund (2017b) p. 177). Under the null hypothesis of independence, the large sample distribution of X_{RS}^2 is $\chi^2_{[(R-1)(C-1)]}$.

117. When the sample size or degrees of freedom is small, adjustments to the X_{RS}^2 test statistic can improve accuracy. These adjustments use an **F-distribution** instead of the chi-square distribution, making the tests more robust for smaller datasets. As mentioned by Steven G. Heeringa, West, and Berglund (2017b), it was Fay (1979), along with Fellegi (1980), who began proposing corrections to Pearson's chi-square statistic based on a generalized design effect. Rao and Scott (1984) later expanded the theory of generalized design effect corrections for these statistical tests, as did Thomas and Rao (1987).

118. The Rao-Scott adjustment requires the calculation of generalized design effects, which are analytically more complex than Fellegi's approach. Nevertheless, Rao-Scott adjusted statistics are now the standard for analysing categorical survey data in software systems such as R, Stata, and SAS.

9.5.4. Tests for group comparisons

119. Comparing group means is a common goal in household survey analysis. For example, researchers might ask: "Is there a significant difference in average income between male and female headed households?" To answer such questions, statistical tests are used, adapted to account for the complexities of survey data, such as stratification, clustering, and unequal selection probabilities. This section explains the methods for testing differences in means, adjusted for survey design, with examples to illustrate their application.

120. To determine whether the means of two groups are significantly different we will introduce t-test and contrasts adjusted for the sampling design.

9.5.4.1. Hypothesis Test for the Difference of Means

121. A hypothesis test is a statistical procedure used to evaluate evidence in favour of or against a statement or assumption about a population. In this process, a null hypothesis (H_0) is proposed, representing the initial statement that needs to be tested, and an alternative hypothesis (H_1), which is the statement opposing the null hypothesis. These statements may be based on some belief or past experience and will be tested using the evidence gathered from the survey data. If it is suspected that the parameter θ is equal to a particular value θ_0, the possible combinations of hypotheses that can be tested are:

122. One of the two hypotheses will be considered true only if the statistical evidence, which is obtained from the sample, supports it. The decision about which hypothesis is true is based on the statistical evidence gathered from the data. This process is called **hypothesis testing**.

123. In many cases, important parameters of interest, such as differences in means or weighted sums of means, can be expressed as a linear combination of various descriptive statistics. These combinations are often used in constructing economic indices or comparing population means. The variance of these combinations is important for understanding the precision of the estimate. That is, parameters can be expressed as a linear combination of measures of interest. The most common cases are differences in means, weighted sums of means used to construct economic indices, etc. Thus, consider a function that is a linear combination of J descriptive statistics, as shown below:

$$f(\theta_1, \dots, \theta_J) = \sum_{j=1}^{J} a_j \, \theta_j \tag{9-31}$$

where the a_j are *known* constants. An estimator of this function is given by:

$$\hat{f}(\hat{\theta}_1, \dots, \hat{\theta}_J) = \sum_{j=1}^{J} a_j \, \hat{\theta}_j \tag{9-32}$$

and its variance is calculated as follows:

$$Var\left(\sum_{j=1}^{J} a_j \, \hat{\theta}_j\right) = \sum_{j=1}^{J} a_j^2 \; Var(\hat{\theta}_j) + 2 \times \sum_{j=1}^{J-1} \sum_{k>j}^{J} a_j \, a_k \, Cov(\hat{\theta}_j, \hat{\theta}_k) \tag{9-33}$$

As seen in the variance expression for the estimator, it requires the variances of the individual estimators, as well as the covariances of pairs of estimators.

124. In the context of comparing means between two populations, there are several potential hypotheses that can be tested. On the one hand, the null hypothesis may state that the means of two populations are equal. On the other, the alternative hypothesis could suggest that the means are different, or that one is greater than or less than the other.

125. Of particular interest is analysing the difference in population means. In order to formulate the hypothesis tests for this case, we need to consider a *superpopulation model*. We assume that y_{hik} correspond to observations from identically distributed random variables Y having means $\mu_{y,j}$ if unit k belongs to domain j, with $j = 1,2$. Then we can define the difference in population means between domains

1 and 2 as $\mu_{y,1} - \mu_{y,2}$. As an example, consider that $\mu_{y,1}$ is the average household income for households with male heads of household, and $\mu_{y,2}$ is the average household income for households with female heads.

126. This difference in means can be consistently estimated by:

$$\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2 \tag{9-34}$$

where $\hat{\bar{Y}}_j$ is the sample estimator of $\mu_{y,j}$ $(j = 1,2)$.

127. Considering the parameter of interest in this section, the hypotheses to test are typically:

- Null hypothesis: There is no difference between the means.
- Alternative hypothesis: There is a difference, which could be in either direction (greater or less).

128. To test one of these hypotheses, the following test statistic is used, which follows a t-student distribution with df degrees of freedom, calculated as the difference between the number of PSUs m in the sample and the number of strata H.

$$t = \frac{\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2}{se\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2\right)} \sim t_{[df]} \tag{9-35}$$

where:

$$\widehat{se}\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2\right) = \sqrt{\widehat{Var}\left(\hat{\bar{Y}}_1\right) + \widehat{Var}\left(\hat{\bar{Y}}_2\right) - 2\,\widehat{Cov}\left(\hat{\bar{Y}}_1\,;\hat{\bar{Y}}_2\right)} \tag{9-36}$$

129. If a confidence interval is needed for the difference in means, it is constructed using the estimated difference and the standard error, along with the appropriate critical value from the t-distribution. This interval provides a range of plausible values for the true difference in means, offering a more complete understanding of the data. It would be constructed as follows:

$$\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2 \pm t_{[df]}\,\widehat{se}\left(\hat{\bar{Y}}_1 - \hat{\bar{Y}}_2\right) \tag{9-37}$$

9.5.5. NSO – Practical example

130. [[In this part an NSO will share its experiences on dealing with statistical comparisons among groups and how do they present the results in tables.]]

9.6. Regression: Modelling survey data

131. Regression modelling is a powerful tool for analysing relationships between variables in survey data. It allows researchers to estimate how one or more independent variables (predictors) influence a dependent variable (outcome). For instance, consider a researcher who is modelling household income (dependent variable) as a function of education level and employment status (independent variables) using household survey data. Such data are often obtained from surveys that adopted complex sampling designs that include stratification, clustering, and unequal probabilities of selection.

132. In this section, we explore how survey weights and sampling design features are incorporated into regression model specification and fitting. We also discuss a parsimonious solution to the challenges posed by weighting. Modelling survey data requires careful consideration of the sampling design to ensure valid

inferences. Incorporating survey weights and adjusting for clustering and stratification allows researchers to produce accurate, representative, and reliable results.

9.6.1. To weight or not to weight?

133. When performing regression analysis on survey data, a key question arises: should we include survey weights in the estimation of regression parameters and their associated standard errors? This question has sparked debate among researchers - Skinner, Holt, and Smith (1989), Pfeffermann (2011) - as there are trade-offs between accounting for the complex design features and simplifying the model for ease of interpretation and efficiency.

134. On the one hand, when including sampling features, we are making sure to achieve **Population Representativeness**, because survey weights ensure that the regression model reflects the true population distribution, correcting for oversampling or undersampling of certain groups. Also, we will obtain **accurate variance estimates**, because we are adjusting for stratification, clustering, and unequal selection probabilities, providing valid standard errors and confidence intervals.

135. On the other hand, including sampling design features may yield to an **increasing of variance**, and can inflate the variance of parameter estimates, particularly if the weights vary widely. Also, extreme or highly variable weights can lead to unstable estimates, where certain observations disproportionately influence the model, making it unstable. For explanatory or analytical purposes (e.g., understanding relationships between variables), unweighted models can sometimes provide more efficient and stable estimates.

136. To answer the question: when to weight?, we can distinguish two scenarios:

- **Descriptive Inference**: Always weight. The primary goal is to reflect the population, and survey weights are essential for accuracy.
- **Analytical Inference**: Consider unweighted or weight-adjusted models. If the goal is to explore relationships or test hypotheses, weighting may not always be necessary, particularly if the model structure includes key design variables (e.g., strata or clusters).

9.6.2. Some inferential approaches to modelling data

137. When working with survey data, one of the key challenges is understanding and addressing the variability inherent in the data. This variability comes from two primary sources. The first source is the **sampling design**, which refers to the way the data was collected. The second source of variability arises from the **model itself**, which is used to analyse the data and make inferences about the population.

138. To combine these two sources of variability into a coherent framework, advanced inferential methods are required. These methods aim to respect the structure of the survey design while also accounting for the assumptions and uncertainties within the model. Two main approaches used for this purpose are **pseudo-likelihood** and **combined inference** ([[Author note: citation needed]]).

139. The **pseudo-likelihood approach** modifies the traditional likelihood methods used in statistical modelling to account for the complexities of the survey design. In simpler terms, it adjusts the standard modelling techniques to ensure that they properly incorporate the way the sample was drawn. This adjustment is crucial because ignoring the sampling design can lead to biased estimates and incorrect conclusions about the population.

140. On the other hand, **combined inference** seeks to integrate the information from the survey design and the model in a unified way. This approach ensures that the uncertainties from both sources —sampling and model— are reflected in the final results. By blending these components, combined inference provides a more comprehensive view of the variability and helps produce more reliable estimates.

9.6.3. Linear models

141. A regression model seeks to explain how changes in one or more independent (explanatory) variables affect a dependent (response) variable. In its simplest form, linear regression examines the relationship between a single independent variable and a dependent variable. The dependent variable is the outcome of interest, while the independent variable represents factors that may influence it. The model also includes an error term, which captures unexplained variability in the data.

9.6.3.1. Basic definitions

142. A simple linear regression model is defined as $y = \beta_0 + \beta_1 x + \varepsilon$, where y represents the dependent variable, x is the independent variable, and β_0 and β_1 are the model parameters. The variable ε is known as the *random error* of the model.

143. For more complex situations, multiple linear regression models allow for the inclusion of several independent variables. This approach helps to account for the simultaneous effects of multiple factors on the outcome. In these models, each independent variable is associated with a coefficient, which indicates the strength and direction of its relationship with the dependent variable. Notice that a positive coefficient suggests that as the corresponding independent variable increases, the dependent variable also increases. Generalizing the previous model, multiple linear regression models are defined by allowing the dependent variable to interact with two or more variables, as presented below:

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon = \sum_{j=0}^{p} \beta_j x_j + \varepsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{9-38}$$

Another way to write the multiple regression model is:

$$y_k = \mathbf{x}_k\boldsymbol{\beta} + \varepsilon_k \tag{9-39}$$

where, $\mathbf{x}_k = \left(1\ , x_{1k}\ , \ldots\ , x_{pk}\right)$ and $\boldsymbol{\beta} = \left(\beta_0\ ,\ , \beta_1\ , \ldots\ , \beta_p\right)^T$.

144. The subscript k refers to the sample element or respondent in the dataset. Regression models are built on several assumptions about the data. Steven G. Heeringa, West, and Berglund (2017a) present some considerations for regression models, which are described below:

- $E(\varepsilon_k \mid x_k) = 0$. That is, the average error for any given value of the independent variable is assumed to be zero, meaning that the model does not systematically over- or under-predict outcomes.
- $Var(\varepsilon_k \mid x_k) = \sigma^2_{y|x}$. That is, the variability of the errors should be constant across all levels of the independent variables, a property known as homoscedasticity (homogeneity of variance).
- $\varepsilon_k \mid x_k \sim N\left(0,\ \sigma^2_{y|x}\right)$ (normality of errors), meaning that the residuals conditioned on the covariates follow a normal distribution. This property also extends to the response variable y_k.

- $cov(\varepsilon_k, \varepsilon_l \mid x_k, x_l) = 0 \; \forall k \neq l$. That is, the errors for different observations should be independent, meaning that the outcome for one observation does not influence another. This way, the residuals in different observed units should not be correlated with the values given by their predictor variables.

145. When these assumptions are met, regression models can provide accurate and unbiased estimates of relationships between variables. The predicted values from the model represent the expected outcomes based on the observed values of the independent variables. This makes regression a useful tool for understanding patterns in data and making informed predictions.

146. Once the linear regression model and its assumptions are defined, it can be deduced that the best unbiased linear estimator is defined as the expected value of the dependent variable conditioned on the independent variables x, as:

$$E(y \mid x) = \hat{\beta}_0 + \widehat{\beta_1} x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{9-40}$$

147. As noted by Steven G. Heeringa, West, and Berglund (2017a), the first authors to empirically discuss the impact of complex sampling designs on regression model inferences were Kish and Frankel (1974), who highlighted the challenges posed by complex sampling designs. Later, Fuller (1975) developed a variance estimator for regression model parameters based on Taylor linearization with unequal weighting of observations under stratified and two-stage sampling designs.

148. As is well known, the use of regression model theory requires certain statistical assumptions to be met, which can sometimes be challenging to verify in practice. In this regard, B. V. Shah, Holt, and Folsom (1977) discuss some aspects related to the violation of these assumptions and provide appropriate methods for making inferences about the estimated parameters of linear regression models using survey data.

149. Similarly, David A. Binder (1983a) obtained the sampling distributions of estimators for regression parameters in finite populations and related variance estimators in the context of complex samples. Skinner, Holt, and Smith (1989) studied the properties of variance estimators for regression coefficients under complex sample designs. Later, Fuller (2002) provided a summary of estimation methods for regression models containing information related to complex samples. Finally, Pfeffermann (2011) discussed various approaches to fitting linear regression models to complex survey data, presenting empirical support for the use of the "*q-weighted*" method, which is recommended in this document.

9.6.3.2. Estimation of parameters

150. When working with survey data, the goal of regression analysis is to estimate the relationships between variables *in the population*. If a complete census were available, the calculation of regression parameters would be straightforward because we would have all the data. However, in practice, we work with survey samples, which introduce additional complexities, particularly when these samples are drawn using complex designs.

151. When estimating the parameters of a linear regression model considering that the observed information comes from surveys with complex samples, the standard approach to estimating regression coefficients and their standard errors is altered. The main reason for this change is that data collected through a complex survey generally does not have an identical distribution, and the assumption of independence cannot be maintained since the sample design is constructed with dependencies (as most complex designs include stratification, clustering, unequal selection probabilities, etc.).

152. In this context, when fitting regression models with such datasets, using conventional estimators derived from traditional methods (such as maximum likelihood, for example) will induce bias because these methods assume the sample data are independently and identically distributed and come from a specific probability distribution (binomial, Poisson, exponential, normal, etc.).

153. For illustrative purposes, the estimation of the parameter β_1 and its variance for a simple linear regression will be shown. The extension to multiple regression parameter estimation is algebraically complex and beyond the scope of this book. Below is the estimation of the slope and its variance in a simple linear regression model:

$$\widehat{\beta_1} = \frac{\sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \left(y_{hik} - \widehat{\overline{Y}}\right)\left(x_{hik} - \widehat{\overline{X}}\right)}{\sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} w_{hik} \left(x_{hik} - \widehat{\overline{X}}\right)^2} \tag{9-41}$$

154. Understanding the uncertainty in the estimated parameters is essential for making valid inferences. In regression analysis with complex samples, variance estimation involves calculating measures that reflect the variability introduced by the sampling design. These calculations often rely on weighted sums and include adjustments for dependencies in the data.

155. For multiple regression, the variance of each coefficient is computed while considering its relationship with other coefficients. This results in a variance-covariance matrix, which summarizes the variability and relationships between all the estimated coefficients. The matrix provides a comprehensive view of the precision of the estimates and is a key tool for interpreting regression results. As a generalization, according to Kish and Frankel (1974), the variance estimation of coefficients in a multiple linear regression model requires weighted totals for the squares and cross-products of all combinations of y and $x = \{1, x_1, \ldots, x_p\}$.

9.6.4. Working with weights

156. When analysing data from complex surveys, a critical question arises: how should survey weights be used in regression models? Steven G. Heeringa, West, and Berglund (2017a) addresses the problem of how to correctly weight regression models and whether expansion factors should be used to estimate regression coefficients when working with complex survey data. In this context, two main approaches exist for incorporating weights into regression models when working with complex survey data:

157. First, the **design-based approach** focuses on making inferences about the entire population. Here, survey weights are essential to ensure unbiased estimates of the regression coefficients. These weights account for the survey design, including unequal probabilities of selection. However, this approach has a limitation: it does not protect against model misspecification. If the model does not correctly describe the relationships in the population, the estimates will still be unbiased for the specified model but not necessarily meaningful for the population. If the researcher fits a poorly specified model, unbiased estimates of the regression parameters would be obtained in a model that does not correctly describe the relationships in the finite population.

158. In contrast, the **model-based approach** argues that weights are unnecessary if the model is correctly specified for the sample, and the sampling is non-informative. This approach assumes that the relationships between the variables are well-represented by the model, regardless of the sampling design. Using weights in this case can increase the variability of the estimates, leading to unnecessarily larger standard errors.

159. The choice between these two approaches depends on the context and the sensitivity of the inferences to the inclusion of weights. A practical recommendation is to fit regression models both with and without weights using statistical software and compare the results. If including weights leads to significant changes in the regression coefficients or conclusions, it indicates that either the model may not be correctly specified or the sampling was informative, and therefore weighted estimates should be preferred. On the other hand, if weights only increase the standard errors without altering the coefficients meaningfully, the model is likely well-specified, and weights may not be necessary.

160. To address the challenges of using raw sampling weights, several adjustments have been proposed to balance accuracy and efficiency. Some are listed below:

[1] **Senate Sampling Weights**: This approach scales weights so that the sum of the weights equals the sample size rather than the population size. The goal is to retain representativeness while reducing the variability of weights. This adjustment is particularly useful in large samples where the raw weights are excessively variable. This approach preserves relative differences in weights.

$$w_k^{\text{Senate}} = w_k \times \frac{n}{\sum w_k} \tag{9-42}$$

[2] **Normalized Weights**: These weights are used to rescale the raw weights to sum to one. This adjustment ensures that the overall weight does not inflate variances unnecessarily. This approach is useful when comparing models with different subsets of data or when variance inflation is a concern.

$$w_k^{\text{Normalized}} = \frac{w_k}{\sum w_k} \tag{9-43}$$

[3] **Pfeffermann Model Weights**: This approach incorporates weights into the likelihood function of the regression model, allowing the model to use weights adaptively. This method combines the benefits of weighting and model-based inference. This approach also adjusts for the variability introduced by weights while retaining design-based properties. It is ideal for models requiring both descriptive and analytical inference.

161. The third solution (called the *q-weighted approach*) was proposed by Pfeffermann (2011) who suggested a slightly different specification of the weights. This adjustment modifies the original weights to reflect the relationships in the data more accurately, reducing variability while still accounting for the complex survey design. It also balances the benefits of both design- and population-based approaches. The steps in this approach are as follows:

[1] A regression model is fitted to the original survey weights, using the predictor variables in the primary regression model of interest.
[2] Predicted survey weights are obtained for each case based on the predictor variables.
[3] The original survey weights are divided by these predicted weight values, creating adjusted weights.
[4] These adjusted weights are then used in the final regression model.

162. The q-weighted approach provides a middle ground. It retains the benefits of the design-based approach by incorporating survey weights but reduces the variance typically associated with their use. At the same time, it considers the relationships captured by the model, aligning more closely with the model-

based perspective. This makes it particularly useful for situations where the choice between the two paradigms is unclear or when both perspectives have merit.

9.6.5. Model diagnostics

163. When using statistical models with household survey data, it is essential to evaluate the quality of the models to ensure the validity of the conclusions. This involves a series of checks and analyses that focus on the assumptions and performance of the model. These checks help confirm whether the model adequately describes the data and whether the results can be trusted. Model diagnostics begin with evaluating whether the model fits the data well. This involves analysing several aspects:

- **Model fit**: It is important to determine whether the model provides an adequate fit to the data, i.e. explains a good portion of the variability of the response;
- **Distribution of errors**: Examine whether the errors are normally distributed;
- **Error variance**: Check whether the errors have constant variance;
- **Error independence**: Verify that the errors can be assumed to be uncorrelated;
- **Influential data points**: Identify if any data points have an unusually large influence on the estimated regression model;
- **Outliers**: Detect points that do not follow the general trend of the data, known as outliers.

9.6.5.1. Coefficient of determination

164. The coefficient of determination, also known as the multiple correlation coefficient (R^2), is a common measure of goodness-of-fit in a regression model. This coefficient estimates the proportion of variance in the dependent variable explained by the model and ranges between 0 and 1. A value close to 1 indicates that the model explains a large proportion of that variability, while a value near 0 suggests the opposite. For surveys with complex sampling designs, the weighted estimator of R^2 is given by:

$$\hat{R}^2_\omega = 1 - \frac{\widehat{SSE}_\omega}{\widehat{SST}_\omega} \tag{9-44}$$

where $\widehat{SSE}_\omega$ is the weighted sum of squared errors given by

$$\widehat{SSE}_\omega = \sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} \left(y_{hik} - x_{hik}\hat{\beta}\right)^2 \tag{9-45}$$

and $\widehat{SST}_\omega$ is the total weighted sum of squares given by

$$\widehat{SST}_\omega = \sum_h^H \sum_{i \in s_{1h}} \sum_{k \in s_{hi}} \left(y_{hik} - \hat{\bar{Y}}\right)^2 \tag{9-46}$$

165. This estimator adjusts the R^2 calculation to reflect the characteristics of the sampling design, such as stratification and unequal selection probabilities, ensuring that survey weights are considered when evaluating the goodness-of-fit of the model.

9.6.5.2. Standardized residuals

166. *Residuals* are the differences between observed and predicted values. Analysing residuals is critical for diagnosing whether the model violates key assumptions. For example:

- Residuals should show no specific pattern when plotted against predicted values or independent variables;
- If the residuals exhibit a pattern, this could indicate non-constant variance (heteroscedasticity) or a non-linear relationship.

167. Graphical analysis is often used to detect issues, with plots of residuals against predicted values serving as a common diagnostic tool. A careful study of the residuals should help the researcher conclude whether the fitting process has not violated the assumptions or if, on the contrary, one or more assumptions are not met, requiring a review of the model specification or of the fitting procedure. To analyse the residuals, Pearson residuals (Steven G. Heeringa, West, and Berglund 2017a) are defined as follows:\

$$r_{p_k} = (y_k - \hat{\mu}_k)\sqrt{\frac{w_k}{V(\hat{\mu}_k)}} \tag{9-47}$$

where $\hat{\mu}_k$ is the estimated expected value of y_k under the fitted model, and w_k is the survey weight for unit k in the complex sample dataset. Finally, $V(\hat{\mu}_k)$ is the variance function of the outcome. These residuals are used to perform normality and constant variance analyses.

168. If the assumption of constant variance is not met, the estimators remain unbiased and consistent, but they are no longer efficient. That is, they are no longer the best in the sense that they no longer have the smallest variance among all unbiased estimators. One way to analyse the assumption of constant variance in the errors is through graphical analysis. This is done by plotting the model residuals against $\hat{y}$ or the model residuals against x_k. If these plots reveal any pattern other than a cloud of points with constant spread, it can be concluded that the error variance is not constant.

9.6.5.3. Influential observations

169. Another set of techniques used for model analysis involves examining influential observations. Certain data points can have a disproportionately large impact on the model. These influential points may not necessarily be outliers but could still affect model parameters significantly. An observation is deemed influential if, when removed from the data set, it causes a significant change in the model fit. It is important to note that an influential point may or may not be an outlier. To detect influential observations, it is essential to clarify what type of influence is being sought. For instance, an observation may be influential for parameter estimation but not for error variance estimation. Common techniques for identifying influential observations include:

[1] **Cook's Distance**: Measures the impact of removing a data point on the overall model fit.

[2] $D_f Beta_{(i)}$ **Statistic**: This statistic assesses the effect of removing a data point on individual regression coefficients, and measures the change in the estimated regression coefficient vector when the observation is removed.

[3] $D_f Fits_{(i)}$ **Statistic**: it evaluates the influence of a data point on the overall model fit, and measures the change in the model fit when a particular observation is removed.

9.6.6. Inference on model parameters

170. After confirming that the model fits well and satisfies its assumptions, the next step is to assess whether the independent variables significantly contribute to explaining the dependent variable. This is done by testing the significance of the regression coefficients. If a coefficient is statistically significant, it suggests that the associated variable has a meaningful relationship with the dependent variable.

171. Given the distributional properties of the regression coefficient estimators, a natural test statistic for evaluating the significance of these parameters is based on the t-distribution and is described as follows:

$$t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{df-p} \tag{9-48}$$

where the degrees of freedom (df) for a household survey (complex samples) is given by the number of PSUs m minus the number of strata H and p is the number of predictor variables in the fitted model.

172. This test statistic evaluates the hypotheses $H_0: \beta_j = 0$ versus the alternative $H_1: \beta_j \neq 0$. Similarly, a confidence interval of $(1-\alpha) \times 100\%$ for β_j can be constructed, as follows:

$$\hat{\beta}_j \pm t_{1-\frac{\alpha}{2}, df}\, se(\hat{\beta}_j) \tag{9-49}$$

9.6.6.1. Estimation and prediction

173. According to Neter, Wasserman, and Kutner (1996), linear regression models are essentially used for two purposes. One is to explain the variable of interest in terms of predictors that may be found in surveys, administrative records, censuses, etc. Additionally, they are also used to predict values of the variable under study, either within the range of values collected in the sample or outside of this range. The first purpose has been addressed throughout this chapter, and the second is achieved as follows:

$$\hat{E}(y_k \mid \mathbf{x}_{obs,k}) = \mathbf{x}_{obs,k}\hat{\boldsymbol{\beta}} \tag{9-50}$$

174. The variance of the predicted value is estimated as follows:

$$\widehat{Var}\left(\hat{E}(y_k \mid \mathbf{x}_{obs,k})\right) = \mathbf{x}'_{obs,k}\, \widehat{Cov}(\hat{\beta})\, \mathbf{x}_{obs,k} \tag{9-51}$$

9.7. Tables

175. Tables are a fundamental tool for disseminating statistics from household survey data. They serve to organize and present numerical results efficiently, minimizing the need for lengthy text descriptions. When well-designed, tables enhance clarity and make it easier for users and wider audiences to interpret survey results. It is therefore important to discuss some core principles and ideas to the preparation and production of tables with survey results.

176. Before we enter detailed discussions, it is important to distinguish three main types of tables that can be used for presenting the results of a survey:

- **Presentation Tables**: These tables are designed to highlight key findings and are often included in reports or presentations; they are concise and focus on results that support specific messages or conclusions;
- **Reference Tables**: These tables provide comprehensive details and are aimed at users who need in-depth information; they are typically larger, covering a wide range of variables and subgroups or domains;
- **Long Tables**: These tables are structured for use in databases or data systems; they contain raw or minimally processed data, organized for further analysis or integration with other datasets.

177. Regardless of the type of table, certain principles should guide their design to ensure they are effective and user-friendly. According to Miller (2004), two fundamental principles should always be considered:

- **Principle 1.** Make it easy for your reader to find and understand the numbers presented in your tables, using clear and concise labels for rows and columns, highlighting key results, and avoiding excessive detail that might overwhelm the reader;
- **Principle 2.** Draw the layout and labels of the table in a simple and direct way, helping to focus attention on the results you want to show, using logical and intuitive ordering of rows and columns, grouping related variables or categories together and minimizing clutter by avoiding unnecessary lines, colours, or decorations.

9.7.1. Presentation tables

178. The primary goal of *presentation tables* is to communicate key results clearly and effectively. They are designed to support the accompanying text by organizing data in a way that emphasizes significant patterns, trends, or stories revealed by the survey. These tables should help readers quickly grasp the main findings without being overwhelmed by excessive detail. These are generally small tables, used to highlight certain key results obtained from the survey, to be presented in press releases, executive summaries, scientific articles or reports, or on landing web pages that showcase survey outputs. They are not expected to provide all results on a topic, but rather to highlight key results that should draw the attention of a reader to some of the main stories the data have produced.

179. In presentation tables, the data should be presented concisely, and organized to support the text with the analysis of the corresponding data. They should be designed in such a way to help readers learn about the key results on the topic provided by the survey. Short, well-designed and formatted tables can provide a lot of information that readers can absorb quickly. This applies to tables published in any vehicle: reports, press releases, articles, electronic publications or websites. The example below illustrates the idea.

180. Presentation tables should have rows (and possibly columns) sorted in a way that helps the reader perceive patterns, such as high or low figures. Such tables will often sacrifice detail in exchange for readability and understanding. Numbers should be presented with no more than 3 or 4 digits altogether. If they area population counts, use thousands. If the figures are percentages, use no more than a single decimal digit, or even present only percentages rounded to the nearest integer, if the precision of the estimates do not warrant providing decimals (e.g. margins of error larger than 1%).

9.7.1.1. Example of how to present key findings

181. **Box. 9.1** shows an example of a presentation table designed to highlight key findings on gender representation among different managerial levels and age groups. The accompanying text contextualizes the data, emphasizing the underrepresentation of women in managerial roles across all ages.

182. The table in **Box 9.1** concisely presents the percentage of women in different occupational categories (non-managers, middle managers, and senior managers) across three age groups. It highlights how a well-designed presentation table can summarize key findings and complement textual analysis. By organizing data clearly and emphasizing critical patterns, such tables enhance the readability and impact of survey results.

Box 9.1. Example of presentation table and corresponding text

Among middle and senior managers, women are outnumbered at all ages. The underrepresentation of women was observed in all age groups. Relative to their share among non-managers, women were outnumbered among middle and senior managers. In all age groups, women accounted for about 4 in 10 middle managers and 3 in 10 senior managers.

TABLE: Share of women (%) by age group and occupation.

Age group	Non-managers	Middle managers	Senior managers
25 to 34 years	44.6	40.3	28.4
35 to 44 years	45.7	38.7	31.3
45 to 54 years	48.3	40.5	31.7

NOTE: The category "women" includes women, as well as some non-binary people. Source: Statistics Canada, Census of Population, 2021. https://www150.statcan.gc.ca/n1/pub/36-28-0001/2024010/article/00005-eng.htm

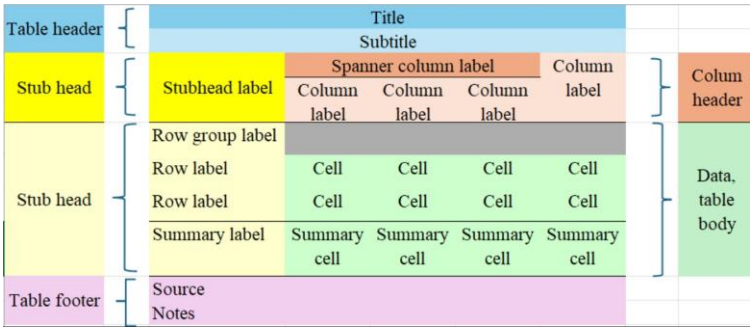
9.7.2. Reference tables

183. Reference tables are longer tables designed to present more comprehensive sets of results from statistical studies. These tables are typically included in reports to provide detailed information, but they should remain manageable in size. A good rule of thumb is to limit them to a maximum of 200 rows and 12 columns. Larger tables should be considered for dissemination as database-like tables, accessible via downloads or interactive websites.

184. Reference tables will typically take core classification, domain definition or *explanatory* variables to define the rows, and have the *outcome* classification or output variables define the columns. In both directions, sorting should typically be such that it is easier for the readers to locate the data that they are most interested in, either using alphabetic sorting or well known classifications. Such tables have in many cases been replaced by access to interactive databases that allow the interested user to obtain the tables they want from a website.

185. Tables (of all types) should be *self-sustaining*. The idea is that each table should have the necessary metadata, so that if copied from one location to another it still makes sense. If you can get your tables to be *self-sustaining*, they will be easier to understand correctly, either in or out of the original context. **Figure 9.3** presents the essential components of a table.

Figure 9.3. Anatomy of a table.



186. The following are the essential components of a well-designed table:

- The *title* (and optional subtitle) of a table is mandatory and must provide a clear and precise indication of the data that will be presented in the table. These elements, combined, must answer the questions about what, where and when regarding the data to be presented inside the table. Be concise and avoid using verbs.

- *Column header* elements should identify the data that is displayed in each column of the table. They must also provide much of the relevant metadata: unit of measurement, time period, geographical area, etc.

- *Row headers* and *stub* elements, provided as the first column in the table, should identify the data that is displayed in each row of the table.

- *Source* of the data must always be provided at the bottom of the table, and must indicate the organization responsible and the name of the survey or study that produced the results contained in the table. The omission of the citation of the source prevents the reader from seeking more information about the data presented, and should be avoided.

- *Notes* are optional, but they can be used to provide additional details about the data as needed to understand and use it correctly. Avoid using long texts, which if needed, would be better placed in a document that is then cited in the Notes section. If there is more than one Note, number sequentially, and use the numbers to indicate the corresponding calls inside the table. Make sure that the calls to Notes are sufficiently distinct from the actual figures / numbers inside the table to avoid confusion.

- *Data* is the most important piece of information that the user expects to get from the table. Therefore, it is essential to present them in a way that is easy to extract the relevant information. For some tables, depending on the message you want to convey, it may be easier to search for information by rows or columns. This should be the most important consideration when deciding whether to present the table in portrait or landscape orientation. Dividing lines, dotted lines, shading, and even spacing can be helpful in guiding the reader to read the table in the 'right' direction.

187. When designing tables to present statistical data, ensuring clarity and consistency is crucial. Start by maintaining uniform spacing across columns to enhance readability, while avoiding unnecessary text or excessive width that can distract from the data. Time series data should always be organized in

chronological order, preferably ascending for reference tables, to provide a clear and logical progression of information. Categorize data using standard classifications to facilitate understanding and comparison across different datasets.

188. The arrangement of rows and columns should follow a clear and logical order, with numerical data aligned to the right to ensure decimal points are neatly aligned. Decimal places should be limited to what is necessary for precision, and rounding should aim for 3–4 significant digits to simplify the data while preserving its integrity. Avoid blank cells, which can cause confusion; instead, use appropriate symbols to indicate missing or "not applicable" values, ensuring the table remains informative and complete.

189. Finally, these practices collectively improve the usability of the table, making it easier for readers to analyse and interpret the data. By adhering to these guidelines, you create a presentation that is both professional and accessible, promoting effective communication of statistical insights.

190. The recommendations provided here to reference tables should also apply to longer tables provided as databases, but these can have additional resources if they are embedded on websites. For example, there may be support for users to sort tables using the values in each column, which would be useful for large tables where the user may be looking for the higher (or lower) values in a given column.

9.7.3. Dissemination of estimates

191. National Statistical Offices routinely produce descriptive statistics, such as totals, averages, proportions, and ratios, based on survey data. These statistics provide valuable insights into key characteristics of the population, such as income levels, employment rates, or access to education. To ensure this information reaches a wide audience, NSOs often use a variety of dissemination channels, including:

- **Public Reports**: Comprehensive reports summarizing key findings from household surveys
- **Online Platforms**: Interactive data visualization tools and downloadable datasets on official websites
- **Press Releases**: Brief summaries of major findings designed to capture public and media attention

192. These dissemination efforts aim to make the data understandable and actionable for policymakers, researchers, and the general public. When publishing tables of results, NSOs strive for clarity and usability. Tables are typically organized to highlight trends, comparisons, and distributions of key variables. Common features of published tables include:

- **Aggregated Data**: Grouping data by domain defining variables like age, sex, region, or socioeconomic status
- **Confidence Intervals**: Including measures of uncertainty to provide context for the estimates
- **Metadata**: Offering detailed explanations of the data collection methods, definitions, and limitations

193. By presenting data in a user-friendly format, NSOs ensure their publications are accessible to a diverse audience. Chapter 10 provides more detailed discussion on presentation of survey findings. However, we note that not all estimates derived from survey data meet the necessary quality standards for publication. Estimates may be suppressed if they are based on small sample sizes, have high variance, or are otherwise unreliable. NSOs use established criteria to determine when suppression is necessary, ensuring that the released data maintains its credibility. To address this issue we can use the following approaches:

[1] **Quality thresholds**: NSOs set predefined thresholds for measures like the coefficient of variation (CV) or standard errors;
[2] **Flagging and suppression**: Estimates that fall below these thresholds are either flagged with warnings about their reliability or omitted entirely from published tables;
[3] **Transparency**: NSOs provide clear documentation explaining why certain estimates are suppressed, maintaining transparency, and trust.

9.8. Data visualization

194. In this section we discuss how to present data and estimates resulting from household surveys using graphics. Effective graphs can reveal patterns, trends, and relationships in the data, making it easier to interpret findings and communicate them to diverse audiences. While standard plots can still be used to show distributions and associations from the raw (unweighted) sample data, these can be misleading for the corresponding population distributions and associations. Therefore it is recommended that modified plots that account for survey weights be used instead.

195. For example, a bar chart showing income distribution should incorporate weights to properly represent the income distribution for the entire population. Similarly, scatter plots exploring associations between variables should use weighted markers or density adjustments to ensure the relationships are accurately depicted. In addition, regarding the display of survey estimates, which are subject to sampling error, it is important to convey this message by presenting not only point estimates, but also standard errors or confidence intervals.

196. When presenting survey estimates, it is essential to recognize that these estimates are subject to sampling error. To effectively communicate this uncertainty, graphs should include measures such as standard errors or confidence intervals. For instance:

- Confidence intervals can be added to bar charts or line graphs to show the range of plausible values for an estimate;
- Error bars in scatter plots can illustrate the variability associated with specific data points.

197. Incorporating these elements into visualizations helps ensure that viewers understand the inherent uncertainty in the survey estimates, fostering more informed interpretations. When the survey units have different sampling weights, these should be taken into account when preparing graphs with their data. The main reason is that weights can be interpreted as the number of population units that each sample unit represents. Hence, it is evident that unequal weights need to be considered in the elaboration of graphs based on such sample data.

198. When graphs are created *without* considering weights, the visual representation reflects the sample characteristics rather than the population. This discrepancy can distort distributions, proportions, or relationships between variables. Incorporating weights ensures that the graphs provide a more accurate representation of the population.

9.8.1. Bar charts

199. When the data of interest are categorical, their descriptive analysis will be done using contingency tables. Bar charts are commonly used to visualize categorical data. For survey data, descriptive analysis of categorical variables typically begins with contingency tables that summarize weighted counts or proportions. These tables can then be used to create bar charts, ensuring the results reflect population-level characteristics rather than just sample data. Ideally one should also aim to display error lines overlaying bars to indicate their respective confidence interval widths, thus conveying the uncertainty of the

corresponding point estimates. Obtaining the weighted counts or proportions and their confidence intervals can be easily done using tools from several software packages, e.g., the survey package in R.

200. As an example, the bar chart in **Figure 9.4** presents a comparison of the number of individuals (Nd) between rural and urban zones, with error lines indicating the confidence intervals for each estimate; it is based off of the data in **Table 9.1**. According to the values in the table, the urban zone shows a slightly higher Nd value than the rural zone, with 78,164 individuals in the urban area compared to 72,102 in the rural area. This difference suggests a higher concentration of people in the urban zone.

Figure 9.4. Distribution of population by area.

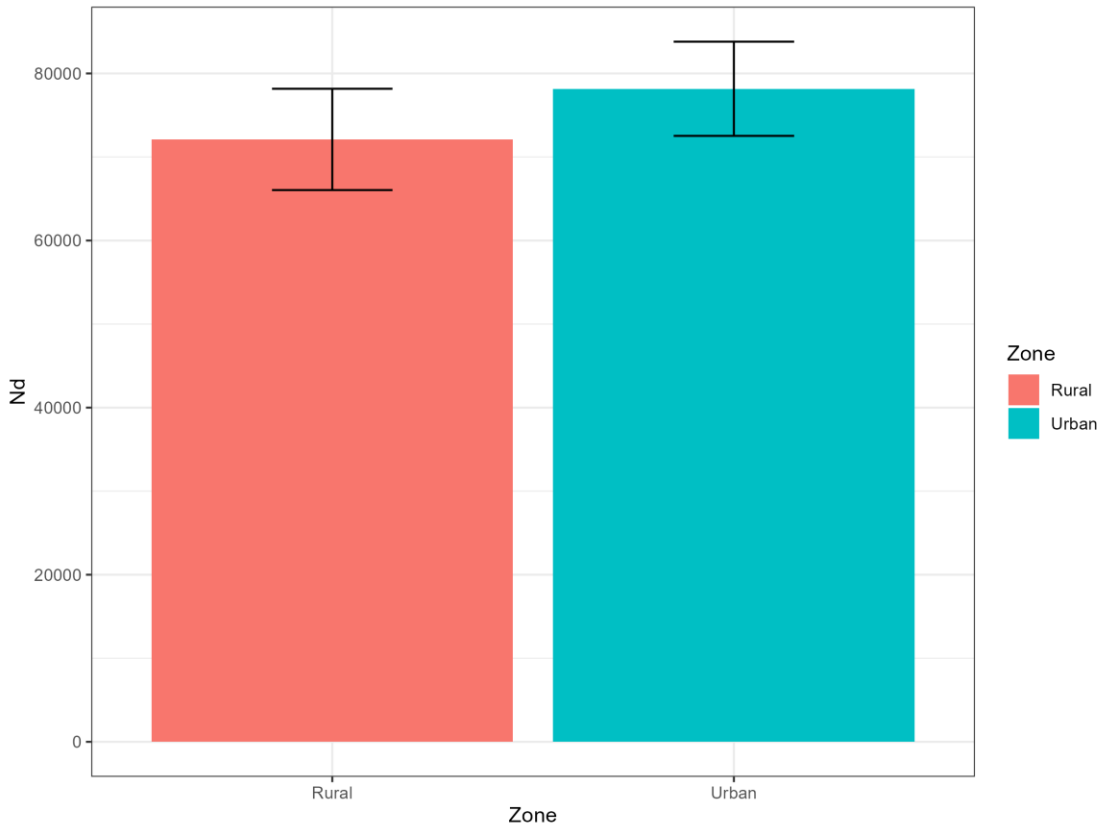


Table 9.1. Distribution of population by area.

Zone	Number of individuals (Nd)	Standard error (Nd_se)	Lower limit (Nd_low)	Upper limit (Nd_upp)
Rural	72,102	3,062	66,039	78,165
Urban	78,164	2,847	72,526	83,802

201. The confidence intervals allow us to assess the precision of these estimates. In the rural zone, the confidence interval ranges from 66,039 to 78,165 individuals, while in the urban zone, the confidence range goes from 72,526 to 83,802 individuals. This overlap between the intervals indicates that, although the urban zone has a higher number of individuals, the difference is not pronounced enough to be statistically significant.

202. Furthermore, the standard deviation of Nd is 3,062 for the rural zone and 2,847 for the urban zone, reflecting similar variability in both zones. This suggests that the estimates are consistent in terms of relative uncertainty, without major differences in data dispersion between the zones.

9.8.2. Histograms

203. Histograms serve to present the distribution of a single numeric (continuous) survey variable or response. If one had a census, then the histogram is a powerful tool to describe the underlying distribution, even for very large datasets. When displaying sample data, however, the sampling weights must be taken into account when estimating frequencies or relative frequencies of population units having values in the specified histogram bins. Modern survey analysis tools can easily provide weighted histograms where the sampling weights are incorporated.

204. Histograms are often seen as precursors to density function estimates. A density estimate can be thought of as a histogram with a large number of bins, providing a smoother view of the data distribution. The survey package in R includes functionality for plotting smoothed density estimates that account for sampling weights, offering a more detailed representation of the population.

205. A common example of visualization in this type of analysis is the use of histograms to represent the distribution of variables such as income. These charts allow us to observe the distribution of the variable of interest in the expanded population and to understand its shape, dispersion, and general trends.

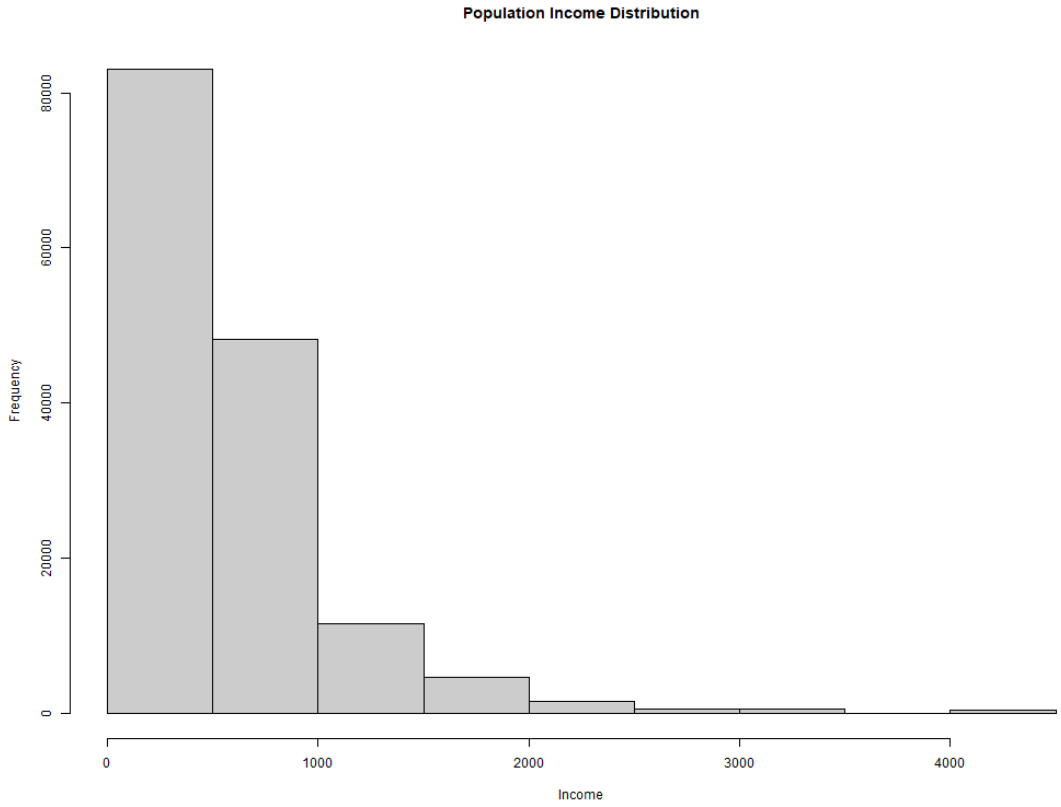
206. It is also common to perform graphical analyses broken down by subgroups, such as geographic areas (urban and rural) or thematic characteristics like sex (male and female). This approach helps identify key differences among specific population subgroups, for instance, by examining income distribution in men and women over the age of 18. Such breakdowns help visualize and communicate potential gaps between subgroups of interest.

207. In this way, charts help to communicate results in an accessible manner, offering a clear and straightforward visual representation for audiences who may not be familiar with the technical details of estimation methods.

208. In **Figure 9.5** the horizontal axis (x) represents income levels, spanning from 0 to over 4000 monetary units, while the vertical axis (y) indicates frequency, meaning the number of individuals within each income range.

209. The distribution shows that most of the population is concentrated at lower income levels, with a particularly high frequency near 0. As income levels rise, frequency declines sharply, indicating a right-skewed (positively skewed) distribution with a smaller proportion of people at higher income levels. The light gray bars visually emphasize this concentration at lower incomes, highlighting a significant disparity in the population's income distribution.

Figure 9.5. Distribution of population income.

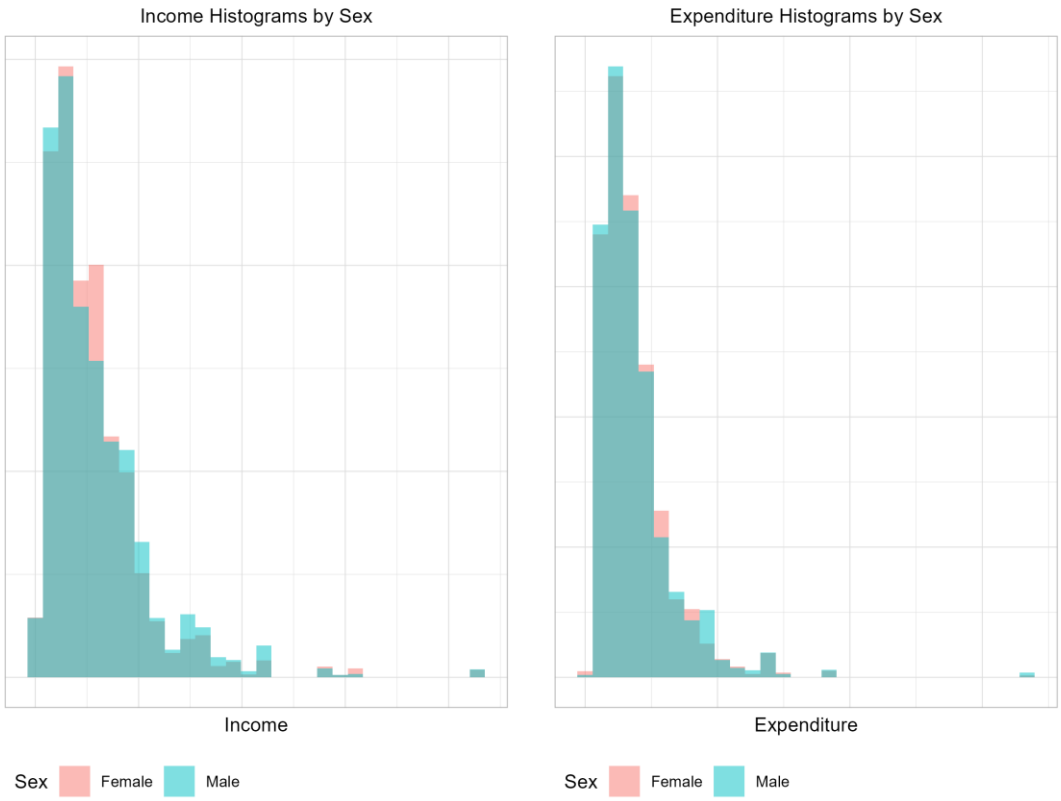


210. As an example, **Figure 9.6** presents two histograms illustrating the distribution of income and expenditure by sex. In the histogram on the left, titled "Income Histograms by Sex," we observe the income distribution, where blue bars represent men and pink bars represent women. The majority of the population, both male and female, is concentrated in the lower income levels, showing a right-skewed distribution. In the lower income levels, there are more men than women, while at higher income levels, the difference is less pronounced.

211. In the histogram on the right, titled "Expenditure Histograms by Sex," the distribution of expenditure is shown, also broken down by sex. Similar to income, most of the population of both sexes is concentrated in the lower expenditure levels, with a right-skewed trend. There is also a higher proportion of men in the lower expenditure levels, while at higher levels, the representation between sexes is more balanced. These histograms exemplify the similarity in the income and expenditure distributions between men and women, although men appear to be slightly more represented in the lower levels of both variables.

212. Histograms, especially when weighted for survey design, are invaluable for exploring and presenting the distribution of continuous variables. Subgroup analyses further enhance their utility, enabling the identification of disparities and trends across different population segments. Combined with smoothed density estimates, histograms provide a comprehensive and accurate view of the population's numeric variables.

Figure 9.6. Histograms of income and expenditure by sex.



9.8.3. Scatter plots

213. Scatter plots are the tool of choice to explore relationships between two continuous variables, potentially revealing patterns or trends in the data. These plots face the two challenges discussed above. First one needs to try and convey in the plot that the different sample observations carry different weights. For small to moderate sample sizes this can be done by plotting circles or dots of varying sizes where the symbol size represents the corresponding observation sampling weight. Plots like these can be obtained using standard bubble plot tools or the scatter plot available in the survey package in R. As stated by Lumley (2010), when dealing with large datasets, displaying all the data points in a scatter plot can be overwhelming and cluttered. Several strategies can help address this issue:

[1] **Subsampling**: Select a smaller, manageable subsample from the full dataset. The subsample should be selected with probabilities proportional to the sampling weights, ensuring that it behaves approximately like a simple random sample from the population. The resulting scatter plot maintains representativeness while being easier to interpret. The subsample obtained in this way behaves approximately as a simple random sample from the survey population.

[2] **Hexagonal Binned Scatter Plots**: Divide the plot area into a grid of hexagons. Instead of plotting individual points, represent each hexagon with shading or size based on the total sampling weights of the points within that hexagon. This approach condenses the data into a clear and interpretable

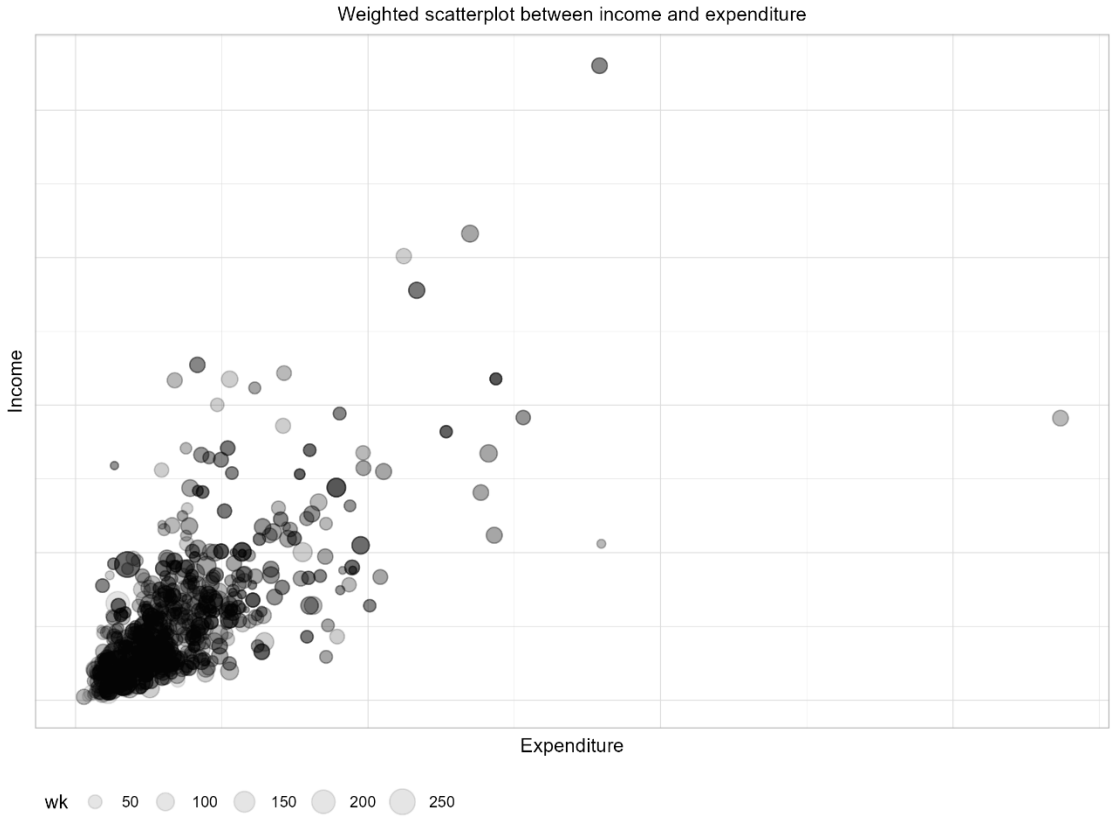
visualization. With complex household survey data, the number of points in a hexagonal bin should be replaced by the sum of the weights for points in the bin.

[3] **Smoothed Scatter Plots**: Avoid plotting individual points altogether and instead estimate and display trends. For example, calculate specific quantiles (e.g., quartiles) of the y-axis variable conditioned on the x-axis variable and smooth these values across the range of the x-axis. This approach highlights trends while minimizing visual clutter.

214. **Figure 9.7** illustrates the weighted relationship between income and expenditure in a population. In this plot, the size of the points represents the weight assigned to each observation. A high concentration of points is observed at lower income and expenditure levels, suggesting that most of the population has low income and low expenditure.

215. Although there is an upward trend, indicating that income and expenditure tend to increase together, the dispersion of points reveals that higher expenditure is not always associated with proportionally higher income. Some larger points, corresponding to observations with greater weight, are distributed across different levels of income and expenditure without concentrating in a single area. Additionally, a few isolated points at high expenditure levels may represent outliers with considerably higher-than-average expenditure. Overall, this plot suggests a positive relationship between income and expenditure, accompanied by significant variability and some exceptional cases.

Figure 9.7. Weighted scatterplot between income and expenditure.



216. Scatter plots are a versatile and effective way to explore relationships between variables in survey data. By incorporating sampling weights and adopting strategies to manage large datasets, they can provide clear, meaningful insights into population-level patterns. Whether using weighted points, hexagonal binning, or smoothing techniques, scatter plots remain a cornerstone of data visualization for continuous variables.

9.8.4. NSO – Practical example

214. [[In this subsection we will include the experience of a NSO on displaying information through graphics.]]

9.9. References

Binder, David A. 1983a. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51: 279–92.

Binder, David A. 1983b. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51 (3): 279–92. https://doi.org/10.2307/1402588.

Binder, David A., and Milojica S. Kovacevic. 1995. "Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach." *Survey Methodology* 21 (2): 137–45.

Dean, Natalie, and Marcello Pagano. 2015. "Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys." *Journal of Survey Statistics and Methodology* 3 (4): 484–503. https://doi.org/10.1093/jssam/smv024.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7 (1): 1–26.

Fay, R. E. 1979. "On Adjusting the Pearson Chi-Square Statistic for Clustered Sampling." *ASA Proceedings of the Social Statistics Section*, 402–8.

Fellegi, Ivan P. 1980. "Approximate Joint Estimation of the Parameters of Multinomial Distributions in the Analysis of Data from Complex Surveys." *Journal of the American Statistical Association* 75 (370): 261–68.

Fuller, Wayne A. 1975. "Regression Analysis for Sample Survey." *Sankyha, Series C* 37: 117–32.

———. 2002. "Regression Estimation for Survey Samples (with Discussion)." *Survey Methodology* 28 (1): 5–23.

Gutiérrez, Hugo Andrés. 2015. *TeachingSampling: Selection of Samples and Parameter Estimation in Finite Population*. https://CRAN.R-project.org/package=TeachingSampling.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. 1953. *Sample Survey Methods and Theory*. Vol. 1 and 2. New York: John Wiley; Sons.

Heeringa, Steven G., Brady T. West, and Patricia A. Berglund. 2017a. *Applied Survey Data Analysis*. Chapman and Hall CRC Statistics in the Social and Behavioral Sciences Series. CRC Press.

Heeringa, Steven G, Brady T West, and Patricia A Berglund. 2017b. *Applied Survey Data Analysis, Second Edition. Statistics in the Social and Behavioral Sciences*. 2nd edition. Chapman; Hall - CRC.

IBM. 2017. *IBM SPSS Complex Samples*. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/23.0/en/client/Manuals/IBM_SPSS_Complex_Samples.pdf.

Jacob, Guilherme, Anthony Damico, and Djalma Pessoa. 2024. *Poverty and Inequality with Complex Survey Data*. https://www.convey-r.org/.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Kish, Leslie, and Martin R Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society, Series B* 36: 1–37.

Kovar, J. G., J. N. K. Rao, and C. F. J. Wu. 1988. "Bootstrap and Other Methods to Measure Errors in Survey Estimates." *Canadian Journal of Statistics* 16 (Suppl.): 25–45.

Langel, Matti, and Yves Tillé. 2013. "Variance Estimation of the Gini Index: Revisiting a Result Several Times Published: Variance Estimation of the Gini Index." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (2): 521–40. https://doi.org/10.1111/j.1467-985X.2012.01048.x.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using r*. *Wiley Series in Survey Methodology*. John Wiley; Sons.

———. 2016. "Survey: Analysis of Complex Survey Samples."

Miller, Jane E. 2004. *The Chicago Guide to Writing about Numbers*. Chicago: University of Chicago Press.

Nations, United. 2005. *Household Surveys in Developing and Transition Countries*. New York, NY: United Nations.

———. 2008. *Designing Household Survey Samples: Practical Guidelines*. Studies in Methods / United Nations, Department of Economic and Social Affairs, Statistics Division Series f. New York, NY: United Nations.

Neter, John, William Wasserman, and Michael H. Kutner. 1996. *Applied Linear Statistical Models*. McGraw-Hill.

Osier, Guillaume. 2009. "Variance Estimation for Complex Indicators of Poverty and Inequality." *Journal of the European Survey Research Association* 3 (3): 167–95. http://ojs.ub.uni-konstanz.de/srm/article/view/369.

Park, Inho, Marianne Winglee, Jay Clark, Keith Rust, Andrea Sedlak, and David Morganstein. 2003. "Design Effects and Survey Planning." *Proceedings of the 2003 Joint Statistical Meetings - Section on Survey Research Methods*, 8.

Pfeffermann, Danny. 2011. "Modelling of Complex Survey Data: Why Model? Why Is It a Problem? How Can We Approach It?" *Survey Methodology* 37 (2): 115–36.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rao, J. N. K., and A. J. Scott. 1984. "On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data." *The Annals of Statistics* 12: 46–60.

Rao, J. N. K., C F J Wu, and K. Yue. 1992. “Some Recent Work on Resampling Methods for Complex Surveys.” *Survey Methodology* 18: 209–17.

Rojas, Hugo Andres Gutierrez. 2020. *Samplesize4surveys: Sample Size Calculations for Complex Surveys*.

Rust, Keith F., Valerie Hsu, and Westat. 2007. “Confidence Intervals for Statistics for Categorical Variables from Complex Samples.” In. https://api.semanticscholar.org/CorpusID:195852485.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SAS. 2010. *SAS/STAT 9.22 User’s Guide - Survey Sampling and Analysis Procedures*. https://support.sas.com/documentation/cdl/en/statugsurveysamp/63778/PDF/default/statugsurveysamp.pdf.

Shah, B. V., M. M. Holt, and R. F. Folsom. 1977. “Inference about Regression Models from Sample Survey Data.” *Bulletin of the International Statistical Institute* 41 (3): 43–57.

Shah, Babubhai. V, Ralph E Folsom, Lisa LaVange, Sara C Wheeless, Kerrie E Boyle, and Rick L Williams. 1993. “Statistical Methods and Mathematical Algorithms Used in SUDAAN.” Research Triangle Institute.

Skinner, Chris J, Daniell Holt, and Tom M F Smith. 1989. *Analysis of Complex Surveys*. New York: John Wiley; Sons.

STATA. 2017. *STATA Survey Data*. https://www.stata.com/manuals13/svy.pdf.

Thomas, D. R., and J. N. K. Rao. 1987. “Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling.” *Journal of the American Statistical Association* 82: 630–36.

Tillé, Yves, and Alina Matei. 2016. *Sampling: Survey Sampling*. https://CRAN.R-project.org/package=sampling.

Westat. 2007. *WesVar 4.3. Users Guide*. http://users.nber.org/~jroth/chap1.pdf.

This page has been intentionally left blank.

CHAPTER 10
DISSEMINATION

Olivier Dupriez (The World Bank)
François Fonteneau (OECD)

CHAPTER 10
DISSEMINATION

10.1. Overview

10.1.1. Introduction

1. Dissemination, as defined by the Guidance on Modernizing Statistical Legislation (UNECE, 2018), refers to the process of making official statistics, statistical analyses, statistical services, and metadata accessible to users. It is an essential phase in the statistical production cycle, ensuring that the data collected, processed, and analysed reaches the intended audience effectively.

2. For household surveys conducted by government agencies, and in particular those agencies part of the national statistical system, the dissemination of survey results must adhere to the United Nations Fundamental Principles of Official Statistics (United Nations, 2014; see Chapter 1). The first Principle (Relevance, Impartiality, and Equal Access) emphasizes the relevance and the impartiality of official statistics. This principle underscores the role of statistical offices as trusted sources of information, providing accurate and transparent data to the public and policymakers. Principle 3 (Accountability and Transparency) is about ensuring the correct interpretation of the data. Principle 6 (Confidentiality) is particularly relevant for survey data dissemination, as individual data collected through surveys are to be strictly confidential and used exclusively for statistical purposes. Principle 9 (Use of International Standards) is about fostering consistency of the survey results.

3. In a rapidly evolving technological landscape, statistical agencies must keep track with what technology offers to maximize users' satisfaction and offer a diversified suite of products for their survey results. User expectations are constantly evolving. The way they consume statistical products has changed significantly and will continue to change even more rapidly in the era of artificial intelligence (AI). To stay relevant, statistical offices must offer data products and services in multiple formats that cater to the needs and preferences of various categories of users. This includes traditional dissemination channels such as printed reports and news releases, but also extends to digital platforms, social media, and interactive tools. The goal is to make data accessible, understandable, engaging, and easily usable for different categories of users. By delivering results in an attractive and user-friendly format, statistical agencies will maintain the trust of their audience and ensure that the data they produce is recognized and utilized for informed decision-making.

4. Section 10.1 of the Chapter offers guiding principles for effectively sharing survey results. Section 10.2 explains how to generate and disseminate core survey products, while Section 10.3 details the creation of data products using emerging approaches.

10.1.2. Guiding principles

5. Successful dissemination of survey results extends beyond the mere release of data; it involves enabling users to quickly locate the information they need, comprehend its significance, and confidently use it in their analyses or decision-making processes. This type of dissemination reflects the organization's commitment to making survey findings accessible, relevant, and impactful for a diverse range of users. Adhering to a set of guiding principles ensures that the products generated from the dataset uphold the highest quality standards. Below, these guiding principles are outlined, drawing inspiration from the National Quality Assurance Frameworks, Manual for Official Statistics (United Nations, 2019). These dimensions of data quality may occasionally conflict, requiring trade-offs. For instance, enhancing timeliness might affect accuracy. The following principles are key:

[1] **Relevance**. Effective dissemination of survey results requires that outputs meet the specific needs of diverse user groups, including policymakers, researchers, media, the private sector, and the public. To achieve this goal, it is essential to engage both primary and secondary users during the design and planning stages of dissemination. Tailored products should be developed, such as in-depth analytical reports for specialists, concise policy briefs for decision-makers, and accessible summaries or infographics for non-experts. Survey results should align with current policy discussions and research priorities, ensuring their practical application in decision-making. Relevance involves addressing the most pressing policy questions, while usability refers to the ease with which users can access, interpret, and exploit the data. Usability can be achieved by providing user-friendly platforms, tools for data exploration, and well-organized reports and visualizations that accommodate the needs of varied audiences.

[2] **Accuracy**. Ensuring the accuracy of survey results involves implementing rigorous data collection, processing, and analysis methods to minimize errors. For dissemination, accuracy requires providing users with clear information about the reliability of the estimates, including confidence intervals and error margins, enabling users to make informed judgments about data quality.

[3] **Timeliness**. The value of survey data is often time sensitive, especially when it informs policy decisions or addresses urgent social and economic issues. Disseminating results promptly ensures that the data are still relevant when published, maximizing its impact. Statistical agencies should strive to release data in accordance with pre-announced schedules to ensure that users can rely on timely updates. A well-planned release calendar is crucial for ensuring the timely and effective dissemination of survey products. By outlining the expected release dates for different survey products, such as data files, reports, and visualizations, a release calendar provides a clear roadmap for stakeholders and helps to manage expectations. It also helps to ensure that resources are allocated efficiently and that deadlines are met by the statistical agency and its staff. A release calendar can be used to coordinate the efforts of different teams involved in the dissemination process, including data analysts, communications staff, and external partners including the media.

[4] **Accessibility and inclusiveness**. Survey results must be easily accessible to all users, regardless of their technical expertise or financial resources. This requires offering data in multiple user-friendly formats, such as reports, tables, visualizations, microdata, and interactive tools or dashboards, while utilizing a variety of dissemination channels, including websites, social media, press releases, and traditional media. Accessibility also involves ensuring affordability, with data provided either freely or at a cost that does not hinder use. Additionally, results should be made available in local languages where relevant. To support users with disabilities, online products should comply with web accessibility standards and guidelines published by the Web Accessibility Initiative (WAI) of the World Wide Web Consortium (W3C).[1]

[5] **Clarity, interpretability, usability.** Survey results must be presented in a clear and interpretable manner, ensuring that users from various backgrounds, including non-experts, can understand and derive meaningful insights from the data. Survey results should be accompanied by explanatory notes accompanying tables, visualizations and other data outputs, guides, and user support that make it easier to interpret and use the data.

[1] See https://www.w3.org/WAI/

[6] **Consistency and coherence**. Survey results must be internally consistent, avoiding any conflicting information within the data. Results should also be coherent with other related data sources, such as previous survey results or complementary datasets, enabling users to compare and integrate data from different sources without discrepancies. Ensuring consistency and coherence enhances the quality of long-term analysis and interpretation.

[7] **Credibility and trustworthiness**. Survey results must inspire both credibility and trust. Credibility is built through professional practices and adherence to recognized international standards and code of practice, ensuring transparency in the methods and processes used to generate the survey results. Building credibility includes disseminating detailed metadata and openly communicating any limitations or potential biases in the data. Trustworthiness is further reinforced by publishing results independently, free from political or external influence, and consistently upholding professional standards. The reputation of the statistical agency plays a crucial role, making it essential to maintain clear communication and protect the integrity of the data at all times. In turn, surveys (and censuses) being very visible operations for the media, the private sector or the population, play a big role in creating trust for the statistical agencies.

[8] **Openness and transparency**. Survey results should be disseminated in a way that ensures broad access and encourages widespread use. Openness involves making both the data and the methodologies used to generate the results available, enabling users to understand how the data were collected, processed, and analysed. This openness supports informed usage and facilitates reproducibility in analysis. Transparency extends beyond simply sharing data. A key component of transparency is the dissemination of comprehensive metadata. Metadata should describe the survey design, sampling methods, data collection procedures, and any adjustments or data cleaning processes, providing users with a full understanding of the data's context, strengths, and limitations. To further support openness and accessibility, it is recommended that survey results be published under an open license, such as the Creative Commons CC BY 4.0[2] or an equivalent, allowing users to freely use, share, and build upon the data while ensuring appropriate attribution.

[9] **Privacy protection.** Statistical agencies must implement robust anonymization techniques and adhere to national data protection regulations and international good practices to ensure that individual responses remain confidential. This includes applying privacy safeguards in published microdata, ensuring that no individual or small group of populations can be identified. Communication regarding these practices is essential for fostering public trust, as individuals are more likely to participate in future surveys if they are confident their privacy is fully protected.

[10] **Long-term usability.** Future re-use and re-purposing of survey data should be guaranteed through proper archiving, as new needs for new analysis will emerge as societal issues will evolve with years passing. Survey data keeps its relevance over time.

10.2. How to disseminate

6. After the completion of data editing and analysis, the dissemination phase of a survey involves creating multiple outputs designed to cater to diverse audiences and use cases. The products generated should be tailored to communicate the findings effectively, meet the needs of different user groups, and maximize the utility of the survey data. The products should be disseminated with a proper communication

[2] https://creativecommons.org/licenses/by/4.0/deed.en

and user-engagement strategy. The usage of these products should be monitored, to gather information that will guide the design of future data collection activities.

7. This section describes the core products and how they should be generated. It also provides suggestions for a communication and user-engagement strategy, and for monitoring usage of the products. This section is about the core products. Section 10.3 (Emerging Approaches) provides recommendations for additional products and for more advanced monitoring of their usage.

10.2.1. Core products

8. Once the survey data have been finalized, the organization will generate a variety of products based on the edited dataset. Engage with stakeholders—including policymakers, researchers, and the public—to ensure that the products address their needs and remain relevant. The core products that will typically be produced and disseminated include the following, which are then described in detail in the subsections.

- **Technical survey report**. This document provides a detailed account of the survey methodology, including data collection, editing, and analysis processes. It includes lessons learned and a post-mortem evaluation. While the full evaluation may not be published, selected insights can be shared to help users understand the strengths and limitations of the survey.
- **Tabulations**. Given the extensive number of potential cross-tabulations that can be generated from a survey dataset, a tabulation plan should be created during the design stage of the survey. This plan will prioritize the tables to be produced and included in survey reports and other published outputs. Additionally, consideration should be given to providing users with access to the microdata or an application that allows for real-time, customized tabulation.
- **Indicators**. Key indicators derived from the survey may be published in a database containing indicators from multiple sources, making it easier for users to access and analyse the most important statistics.
- **Charts, maps, and infographics**. Visual representations of the data, such as charts, maps, and infographics, can be produced for both print and online publications. These visuals help simplify complex data.
- **Descriptive and analytical survey report**. This report presents descriptive statistics and an analysis of the survey results, highlighting key findings and their implications for policy or research. It is often the primary reference document for users seeking comprehensive information about the survey outcomes.
- **Briefs, blogs, regional and special reports, and press releases**. To communicate key findings quickly and effectively, shorter, more digestible content can be produced, such as policy briefs, blogs, and press releases. These outputs are aimed at a broad audience, including media outlets, policymakers, and the general public. Regional reports have their own audience, form local governments to non-governmental organizations to parliamentarians.
- **Microdata**. For more advanced users, anonymized microdata can be made available either as public-use files or research-use files, depending on the confidentiality requirements and data access policies.

10.2.1.1. Technical survey report

9. The technical report of a survey plays a vital role in both informing data users and preserving institutional memory. Targeted at a technical audience, it establishes and reinforces the credibility of the data collected. To ensure completeness and accuracy, the report should be developed progressively throughout the survey process rather than completed at the end. Ideally, it should be finalized and disseminated concurrently with the publication of the first survey results. The technical report can be combined with the analytical survey report to provide a comprehensive overview of the survey process and findings.[3]

10. The technical report should include the following key elements, drawing on international good practice and on the structure of the Data Documentation Initiative DDI-Codebook metadata standard for structure and format.[4]

- **Survey objectives and background**. A brief abstract outlining the purpose and significance of the survey.
- **Key contributors**. Identification of primary investigators, sponsors, and other organizations involved in the survey design and implementation.
- **Survey methodology**. Detailed descriptions of the survey methodology and questionnaire design, highlighting the rationale behind choices made.
- **Sample frame and design**. Information regarding the sample frame, sampling design, and the calculation of sample weights, including sample calibration methods employed if any.
- **Data collection details**. Comprehensive insights into the data collection process, including dates, interviewer training, supervision, organization of fieldwork, and any issues encountered.
- **Mode of data capture**. Explanation of the techniques used for data capture during the survey.
- **Data processing and quality control**. An overview of the data processing steps and quality control measures implemented to ensure data integrity, including data editing.
- **Analysis methods and tools**. Description of the analytical methods and tools used to process and interpret the data.
- **Limitations and potential sources of error**. A transparent discussion of limitations, including response rates, possible biases, and other factors affecting data reliability.
- **Appendices**: Supplementary materials and documentation that support the main report (including a copy of the questionnaire).

11. Be transparent about limitations and known weaknesses and provide information on comparability with previous surveys and changes in standards over time.

[3] The standardized structure of Demographic and Health Survey (DHS) reports provides a good example of such well-structured and comprehensive reports. They combine the technical and analytical reports.

[4] See DDI Codebook specification at https://ddialliance.org/Specification/DDI-Codebook/2.5/

10.2.1.2. Tabulations

12. The production of a set of pre-designed tables will provide users with quick access to key relationships and patterns within the data. A tabulation plan should be developed at the design stage of the survey, following consultation with analysts and the main users. By designing a set of core tables in advance, statistical agencies can ensure that the most critical and commonly needed combinations of variables will be readily available to users. Providing these tables in electronic formats such as CSV or MS-Excel enhances accessibility and encourages reuse for further analysis.

13. Given the vast array of possible variable combinations, however, it is impractical to pre-tabulate every conceivable outcome. To address this challenge, agencies may consider implementing a system for custom tabulation on demand, allowing users to generate tables tailored to their specific needs and interests.[5] This flexibility enables more personalized and detailed analyses, enhancing the utility of the survey data. If feasible, an online tabulation system can offer users the convenience of generating custom tables in real time, though such systems can be complex, particularly in terms of incorporating weighting and filtering methods, or managing confidentiality. Careful design is necessary to minimize the risk of privacy breach, misuse or misinterpretation of the data. Another option is to provide legitimate users with access to anonymized microdata, allowing them to generate the tables they need.

14. Regardless of whether fixed or custom tables are employed, it is crucial to implement statistical disclosure control measures. These safeguards protect respondent confidentiality and ensure that sensitive information is not inadvertently disclosed.

10.2.1.3. Database of indicators

15. Indicators are statistical summary measures that provide valuable insights into survey data, facilitating the tracking of progress, informing policy decisions, and supporting research initiatives. They provide a standardized approach to measuring and comparing data across various regions, time periods, and demographic groups. This standardization enables policymakers to assess the effectiveness of programs and initiatives through clear and comparable metrics. By distilling complex datasets into understandable metrics, indicators make it easier for stakeholders to grasp key trends and issues, directing attention toward critical areas requiring action. Moreover, indicators serve as baseline data for further research and analysis, allowing for the tracking of changes over time and facilitating longitudinal studies and trend analysis.

16. Some surveys, such as the Multiple Indicator Cluster Surveys (MICS) conducted by UNICEF, the Demographic and Health Surveys (DHS), and Labor Force Surveys (LFS), are specifically designed to generate pre-defined sets of indicators. Other surveys, including multi-topic household surveys and thematic surveys, while primarily research-focused, can also yield valuable indicators.

17. To enhance their utility, indicators should be organized into comprehensive databases that integrate data from multiple surveys and sources rather than being limited to individual datasets. This approach promotes compatibility and comparability across various data sources by employing standardized definitions and methodologies. Detailed metadata for each indicator is crucial and should include definitions, data sources, methodologies, and collection dates.

[5] Online tabulation tools are frequently utilized for processing population census data, such as through the use of the REDATAM application. Implementing these tools for sample survey data, which necessitate applying sample weights, is a more complex process.

18. Additionally, databases should be accessible via Application Programming Interfaces (APIs), allowing users to programmatically access and integrate indicators into their own systems.[6] Consideration should be given to utilizing standards such as Statistical Data and Metadata eXchange (SDMX)[7] for better interoperability.

19. Provide thorough information on sampling errors and confidence intervals to convey the precision and reliability of the indicators. Comprehensive documentation on the methodologies used to calculate indicators must also be made available to ensure transparency and reproducibility.

20. Last, compliance with national and international standards, as well as the use of internationally recognized definitions, will enhance compatibility across different surveys and data sources. By integrating indicators from various surveys and sources, a holistic view of the data landscape can be provided, ultimately enriching the understanding and application of survey findings.

10.2.1.4. Charts, maps, and other infographics

21. The production of charts, maps, and infographics plays a key role in effectively disseminating the results of household surveys. These visualizations transform complex datasets into easily digestible and visually engaging formats, significantly enhancing comprehension and fostering engagement among diverse audiences. Visual tools help make survey results more accessible and understandable, emphasizing key trends, patterns, and outliers that might be overlooked in textual or tabular data.

22. Charts, maps, and infographics not only capture the audience's attention but also hold their interest more effectively than plain text. Interactive visualizations allow users to explore data in depth, creating a more personalized experience that invites deeper engagement. By conveying messages clearly and concisely, well-designed visualizations minimize the potential for misinterpretation, ensuring that individuals with varying levels of data literacy can grasp the essential insights, thereby broadening the impact of survey results.

23. To ensure accessibility for all users, use colour palettes that accommodate colour-blind individuals (e.g., ColorBrewer[8]) and maintain sufficient contrast between elements for those with visual impairments.

24. Collaborating with data visualization experts can elevate the quality of visual outputs, leveraging modern technologies to create dynamic and impactful representations. Detailed charts and maps should support the narrative and highlight significant findings, while infographics can summarize key points, enabling readers to quickly engage with the content.

25. Dissemination strategies should ensure that visualizations are easily embeddable on websites and social media platforms, expanding their reach to broader audiences. Utilizing modern visualization tools can facilitate the creation of interactive visualizations that allow users to dynamically explore the data. For those without extensive coding experience, some platforms offer user-friendly interfaces that enable the development of sophisticated visualizations.

[6] See for example the StatCompiler application for DHS surveys at https://www.statcompiler.com/en/

[7] https://sdmx.org/

[8] See https://colorbrewer2.org/

26. It is important to acknowledge that distributing maps can be a sensitive matter. When generating maps, it is essential to ensure that the boundary files utilized conform to the organization's established guidelines and regulations.

10.2.1.5. Descriptive and analytical survey reports

27. Although preliminary results can be disseminated under specific conditions—ensuring that the preliminary nature is clearly stated and that core quality control measures have been implemented to mitigate significant biases—the descriptive and analytical reports will usually be the first public output of the survey that delivers a thorough examination of the data. These reports represent the official results of the survey and often serve as the primary output for many users.

28. To ensure broad accessibility, the reports should be available in the primary languages of the country, facilitating understanding among diverse stakeholders. They should include both descriptive statistics and in-depth analyses, supported by graphs and tables that highlight key findings, making the data more comprehensible and actionable for users.

29. The descriptive and analytical reports must strive to inform policy and decisions without political bias, providing an objective analysis grounded in the survey findings. They should also include information on sampling errors and clearly communicate the precision and uncertainty of the results to foster trust in the data.

30. Applying best practices in technical writing, the reports should undergo thorough review and copy-editing to ensure clarity and accuracy. While a printed version may be useful for some users, it is essential to provide a machine-readable format. A PDF version is recommended, and accompanying tables should also be available as electronic annexes (e.g., CSV or Excel) for user convenience. Additionally, implementing a chatbot to facilitate natural language interaction with the documents can enhance user engagement and accessibility (see Section 10.0 of this chapter).

31. Agencies responsible for collecting and processing the data may not always possess the necessary subject matter expertise to produce the most relevant analyses. Therefore, establishing collaborations with academia or other external experts can enhance the quality of the analytical reports. These collaborators will require access to the microdata, which should be provided in a formal setting that includes confidentiality agreements and technical measures to ensure secure access and usage of the data (see Section 10.2.1.7 on microdata dissemination).

32. Timeliness is crucial in the release of the reports. It is not necessary to exhaust all topics before publishing; rather, the main results should be made available promptly, with specialized analyses conducted and disseminated subsequently. This approach ensures that users receive timely insights that can inform decision-making and policy development.

10.2.1.6. Briefs, regional and thematic reports, blogs, press releases

33. Producing a collection of briefs, specialized reports, blogs, and press releases will present survey findings in a narrative form that engages specialized and non-expert audiences. These tailored communications not only make the data more relatable but also highlight specific topics, ensuring that the most relevant and impactful aspects of the survey results are emphasized. By focusing on different audience segments, these materials effectively communicate the survey’s key messages to stakeholders, policymakers, and the general public.

34. Specialized outputs enhance the visibility and relevance of survey results by breaking down complex findings into more concise, accessible formats. Briefs and blogs can provide quick insights that inform

decision-making and stimulate public discourse, while also attracting attention to the main findings and encouraging further exploration of the full survey report. This targeted approach helps maximize the use and value of the survey results, ensuring that they reach those who need the information most.

35. For surveys representative at sub-national levels, publishing regional reports in addition to the main national report offers several benefits. Regional reports provide a more granular view of data, allowing users to identify and analyse trends and disparities at the sub-national level. This is particularly valuable for parliamentarians, decentralized governments, and local non-governmental organizations (NGOs) and communities who need to make informed decisions and allocate resources based on regional specificities. By focusing on regional data, these reports can help to address the potential limitations of national sampling frames and improve representativeness in areas with diverse populations. Additionally, regional reports can enhance engagement with stakeholders at the local level, fostering a better understanding of data needs and promoting evidence-based policymaking. However, it is important to note that publishing regional reports requires careful consideration of sample size and representativeness. To ensure that regional results are reliable, it may be necessary to increase the overall sample size or use stratified sampling techniques to allocate a sufficient number of respondents to each region. Failure to address these issues could lead to misleading or inaccurate regional data, undermining the credibility of the survey and its findings.

36. Press releases play a crucial role in gaining media coverage for survey findings. By using clear and engaging language, a press release can capture the attention of journalists and generate interest in the survey's findings. Additionally, a press release should include contact information for a spokesperson from the statistical agencies who can provide further details or interviews.

37. Collaborating closely with external experts—such as journalists, academic researchers, and specialized organizations—can enhance the quality and effectiveness of these communications. These collaborations ensure that the messages are not only accurate but also crafted in a way that resonates with the intended audience, thereby further enhancing the survey's relevance and impact.

10.2.1.7. Public use files or scientific use microdata

38. Statistical agencies “should have the authority to permit access to microdata for researchers under certain conditions and terms. These conditions should be included in the statistical law […].” (UNECE, 2018) Sharing anonymized and well-documented versions of microdata with legitimate users, under clear conditions, will contribute to fostering research, informing policy development, and enhancing the overall utility and impact of survey data. Furthermore, sharing microdata may be a contractual obligation tied to funding agreements with survey sponsors, making it crucial to clarify these obligations from the outset to avoid misunderstandings and ensure compliance. By providing access to anonymized microdata from national sample surveys, statistical agencies empower subject matter experts to conduct nuanced and detailed analyses, yielding deeper insights that improve the relevance of the survey findings.[9]

39. Microdata accessible to third parties can take the form of public use files (PUF) accessible under light constraints, or scientific use files (SUF) accessible to the research community under more strict conditions. It can also be implemented by operating data enclaves.

40. The dissemination of microdata must strictly adhere to legal and ethical standards to protect respondent privacy and maintain public trust. Implementing formal processes for compliance with data

[9] See Dupriez and Boyko, 2010. Dissemination of Microdata Files - Principles, Procedures and Practices. IHSN Working Paper No 005. Available at http://www.ihsn.org/home/node/120

protection laws and ethical guidelines is vital. Terms of use should be clearly communicated to all users, outlining any restrictions on data usage and requirements for data protection.

41. Proper anonymization of data using statistical disclosure control techniques is a critical component of this process. This includes removing or altering identifying information to prevent the re-identification of individuals, particularly in sensitive data like geographic location data (e.g., GPS coordinates). When releasing geo-identifying information, additional anonymization measures should be taken. Documentation of the anonymization process and education for users about these techniques will foster transparency and trust.

42. Providing detailed metadata for the disseminated microdata is equally important, as it enables users to understand the dataset's context and content. Utilizing metadata standards like the Data Documentation Initiative (DDI) Codebook to document the dataset comprehensively helps ensure clarity.

43. Agencies should clearly distinguish between various versions of the micro-datasets—such as unedited, non-anonymized fully edited, an anonymized public use files. Clear versioning aids users in understanding the level of processing and anonymization applied to each dataset. To ensure accessibility, data should be provided in open formats like CSV, as well as in formats compatible with major software applications commonly used by researchers.

44. Last, facilitating proper citation of the data is essential. Implementing a Digital Object Identifier (DOI) system encourages users to cite the data accurately, ensuring that data creators receive appropriate credit and facilitating tracking of data usage.[10]

10.2.2. Dissemination, communication, and user engagement

45. Dissemination and communication refer to the processes involved in sharing and promoting survey results to a wide audience, ensuring that the information is accessible, understandable, and usable.

10.2.2.1. Dissemination

46. Once a survey product is finalized, it is disseminated through various relevant media, with the organization's website serving as the central hub for all outputs. Adhering to web accessibility standards, such as the Web Content Accessibility Guidelines (WCAG)[11], will ensure that all users, regardless of their abilities, can access the information.

47. Different products may need different media and strategies. In addition to the website, employing diverse media channels—including social media, podcasts, television, and radio broadcasts in local languages—can help reach a broader range of stakeholders.

10.2.2.2. Communication

48. A well-crafted communication strategy is crucial for disseminating survey results effectively and ensuring their utilization. Successful dissemination demands proactive efforts to make sure the findings reach the target audience. This involves promoting data releases via various media channels, newsletters, webinars, and events, along with interacting with users to provide support, training, and address their queries. Such a strategy facilitates continuous engagement between the organization and data users, helping users to fully comprehend and use the survey results.

[10] https://datacite.org/create-dois/

[11] https://www.w3.org/WAI/standards-guidelines/wcag/

49. Communication significantly enhances the effectiveness of survey dissemination. While dissemination ensures data availability, communication guarantees users understand the data, its implications, and how to utilize it efficiently. This process includes engaging users early to understand their needs and expectations, providing updates during data collection, and promoting the use of survey outcomes through targeted outreach and training. Effective communication not only boosts the visibility of survey data but also builds trust and collaboration between data producers and users.

10.2.2.3. Engagement channels

50. User engagement is essential for effective survey results dissemination**.** By maintaining proactive engagement throughout the survey lifecycle, statistical agencies can identify areas for improvement, ensure the collected data are relevant and tailored to the users' needs, and build stronger relationships with stakeholders. To maximize reach, a variety of communication channels should be employed:

- **Traditional media**. Press conferences and collaborations with news outlets can extend reach to the general public.
- **Social media**. Platforms such as Twitter, Facebook, LinkedIn, and YouTube can effectively publicize results and engage a broader audience.
- **Newsletters**. Regular newsletters can keep stakeholders informed about new data releases, reports, and updates.
- **Focus groups, public events, and webinars**. Launching data through events or hosting webinars can facilitate explanations of findings and directly address user questions.
- **Training and workshops**. Offering targeted training sessions can help users better understand and apply the survey results.

51. User engagement should not be a one-way process; providing users with easy avenues for feedback is essential. This can include a simple "Contact Us" feature on a website, as long as inquiries are responded to promptly and used to establish a two-way communication channel between data producers and users.

10.2.3. Monitoring and analysing survey products usage

52. Tracking how survey products are used helps gauge their reception and utility among the target audience. By analysing usage patterns, agencies can identify demand for certain data types and make informed decisions about future releases and improvements. To effectively measure and understand survey product usage, statistical agencies can employ a variety of methods. These include:

- **User surveys and feedback:** Conducting surveys or questionnaires can gather direct feedback from users about their satisfaction and needs. Conducting user surveys is most effective when they are well-designed and strategically timed. It is important to recognize that surveys do not always yield unbiased results, as responses can be influenced by how questions are phrased or by self-selection bias among participants.
- **Web usage analytics:** Tracking website traffic, page views, and time spent on specific pages can provide insights into user behavior and preferences. Tracking the number of downloads of data files and citations of survey reports can measure the impact of the products.
- **Social media analytics:** Monitoring social media engagement, such as likes, shares, and comments, can help assess the impact of survey products.

53. Web usage analytics tools, such as Google Analytics, Matomo, and others, provide easy-to-obtain valuable insights into how users interact with the data on a website. Web usage analytics tools can track:

- **Page views**. The number of times a specific page (e.g., the page hosting the survey results) has been viewed.
- **Unique visitors**. The number of distinct users who access the data, providing insight into the reach of the dissemination effort.
- **Downloads**. The number of times survey datasets, reports, or other resources are downloaded, which helps measure the demand for specific products.
- **User demographics**. Basic information on the users (e.g., their geographical location, device types), which can help tailor future dissemination efforts to target specific user groups.

54. These metrics help quantify user interest and identify which data products are most in demand. Analysing trends over time allows statistical agencies to make informed decisions about which data to promote or improve in their future activities.

55. Setting up a web analytics tool like Google Analytics or Matomo involves installing a tracking code on each page of a website to monitor user interactions and behavior. After the initial setup, it is essential to configure the tool's settings beyond the defaults to gain the most actionable insights. This includes defining custom goals and events, such as tracking specific user actions (e.g., downloads, video plays, or form submissions) that align with the site's objectives. Privacy settings are crucial, as both tools allow for adjustments to comply with regulations like GDPR, including options to anonymize IP addresses and manage cookie consent. Additionally, setting up filters to exclude internal traffic, such as from employees or developers, will prevent data from being skewed. This careful customization ensures the web analytics tool provides meaningful and accurate data.

56. By analysing these data, statistical agencies can gain valuable insights into how their survey products are being used and identify areas for improvement. Agencies can then improve their overall dissemination strategy and their specific products, by considering:

- **Further tailoring content:** Creating customized products, such as data visualizations or interactive dashboards, for specific target audiences.
- **Improving accessibility:** Making survey products more accessible by providing them in multiple formats, etc.
- **Enhancing user experience:** Optimizing the design and layout of printed documents, audio, videos, websites and data portals to improve user experience.
- **Promoting data literacy:** Organizing workshops and training sessions to help users understand and interpret survey data.
- **Collaborating with partner organizations** to promote the use of survey data and to identify new opportunities for collaboration (NGOs, development partners, etc.).

10.2.4. Main challenges

57. Disseminating survey results in a way that maximizes their impact poses several challenges for statistical agencies. While the goal is to make data accessible, relevant, and useful to a broad audience,

agencies must navigate various obstacles to achieve this. Below are some of the key challenges faced during the dissemination process:

[1] **Balancing accessibility and technical complexity**. One of the primary challenges is making complex survey data accessible to a diverse audience. Users range from highly skilled researchers to the general public, each with varying levels of technical expertise. While experts may demand detailed datasets and complex tabulations, less experienced users need simpler, more easily interpretable outputs. Creating products that meet these diverse needs without overwhelming or alienating any group is a delicate balance.

[2] **Meeting diverse user needs**. Different user groups—policymakers, researchers, businesses, journalists, and the public—have varying data needs. Policymakers might want concise, actionable insights, while researchers may require detailed datasets. Journalists might seek compelling infographics, and businesses may demand customized data feeds. Fulfilling these diverse needs without diluting the value of the data or spreading resources too thin is a major challenge for statistical agencies.

[3] **Ensuring timeliness of releases**. Timely dissemination is essential to ensure that survey results are relevant to ongoing policy discussions and public debates. However, delays can occur due to lengthy data processing, validation, and analysis phases, or because of resource constraints within the agency. Overcoming these delays while maintaining data quality and accuracy is a persistent challenge. Leveraging new technologies, including AI solutions, can contribute to ensuring timeliness.

[4] **Resource and capacity constraints**. Many statistical agencies face limitations in terms of funding, technical infrastructure, and skilled personnel. These resource constraints hinder the agency's ability to produce high-quality, user-friendly data products, such as interactive tools, visualizations, or APIs. Developing and maintaining quality, complementary dissemination digital and non-digital products, requires significant investments in technology and staff capacity.

[5] **Adapting to rapid technological change**. The way users consume data is evolving rapidly with new technologies and platforms emerging constantly. Statistical agencies may struggle to keep pace with these changes, as they often rely on more traditional dissemination methods like printed reports or static tables. Transitioning to more modern approaches, such as data dashboards, APIs, and mobile-friendly platforms, can require a significant overhaul of existing systems and workflows.

[6] **Promoting awareness and engagement**. Even with high-quality data products, an agency's dissemination efforts may fall short if users are unaware of the available data or do not engage with it. Effective promotion of survey results requires targeted communication and marketing efforts, which are not traditionally the core competencies of statistical agencies. Ensuring that data products reach the right audiences and are effectively used requires a dedicated strategy for outreach and user engagement, often through social media, newsletters, and events.

[7] **Ensuring consistency.** Ensuring consistency and reconciliation between survey data and other data sources from the same statistical agency is a critical challenge in survey data dissemination. Data discrepancies can occur between specific survey results, and other data products from other sources (administrative sources, censuses, other surveys, modelled data). Many users will struggle to understand the reasons for these discrepancies or inconsistencies, which can undermine the credibility and can lead to confusion and mistrust among users. To address this challenge,

statistical agencies must implement sound data quality management processes, including data validation, cleaning, and reconciliation, in line with the guidance provided in other chapters of this Handbook. Additionally, it is important to document the methods used to collect, process, and analyse the data, and to make this information available to users, so they can understand why the specific survey data could be different than other related data products from the statistical agency.

[8] **Ensuring data quality and trust**. Maintaining high data quality is fundamental but ensuring that users perceive the data as trustworthy is equally important. In an era where misinformation is prevalent, agencies must work harder to build and sustain trust in the data they disseminate. This requires transparent methodologies, clear communication of limitations, and consistent delivery of impartial data. Agencies must actively combat any perceptions of bias or unreliability that could undermine the credibility of the survey results.

[9] **Sustaining long-term relevance**. Even after successful dissemination, statistical agencies must ensure that the data continue to be relevant and useful over time. User needs evolve, and the way data is consumed and analysed changes with new tools and technologies. Agencies face the challenge of continuously adapting their dissemination strategies to remain relevant in the long term, updating data formats, revising outputs, and incorporating user feedback into future data releases.

10.3. Emerging approaches

58. The way agencies share their survey results is a space for innovation. It is crucial to recognize two points: innovation rates differ globally and by sector, and AI will significantly influence future consumption of survey data. As generative AI evolves, agencies must adapt to technological advances and user preferences.

59. Innovative survey data products that may be generated and disseminated include:

- **Programs and scripts**. Disseminating code such as R or Python programs and scripts that replicate the data tabulation or analysis enables secondary users to build on the work and ensure reproducibility.
- **Data as a service**. Application Programming Interfaces (APIs) are sets of rules and protocols that allow different software applications to communicate with each other. They can be used for survey dissemination by providing a way for survey data to be accessed and integrated into other applications, such as data analysis tools or visualization platforms.
- **Scrollytelling**. Scrollytelling is a storytelling technique that combines scrolling motion with visual content and text. It can be used for survey dissemination by creating interactive and engaging presentations of survey data that guide users through the findings in a visually appealing and informative way. Scrollytelling can help to make complex data more accessible and understandable.
- **Interactive tools and dashboards**. Additionally, a tool may be developed to allow users to generate custom tables, enabling more personalized exploration of the data. Platforms or dashboards allowing users to interactively explore the data and generate their own visualizations or analysis.

- **ChatBots**. Chatbots are AI-powered virtual assistants that can interact with users through text or voice. They can be used for survey dissemination by providing personalized assistance to respondents, answering questions about the survey.
- **Podcasts and audio products**. Podcasts are audio broadcasts that can be downloaded or streamed online**.** They can be used for survey dissemination by providing informative and engaging content about the survey's purpose, methodology, and findings. Podcasts can reach a wide audience and help to raise awareness of the survey's importance and relevance. Podcasts and other audio products are very popular and effective ways to disseminate key findings in countries with lower data literacy.
- **Videos**. Short videos can be used to explain key findings or showcase how to use the data tools and resources.

10.3.1. Programs and scripts for reproducible analysis

60. Sharing programs and scripts (i.e. the code) used in survey data processing will add value and credibility to the data by increasing transparency, reproducibility, and facilitating the use of disseminated microdata by third parties.[12] By making the code used to generate survey outputs publicly available, statistical agencies promote transparency and reproducibility, which are foundational principles in scientific research and data analysis. Reproducibility ensures that other researchers can verify findings, thereby strengthening the credibility of the results. By providing microdata along with the necessary tools and guidance for their efficient and responsible use, statistical agencies empower users to engage with data meaningfully and effectively.

61. Sharing scripts and programs facilitates the reuse and repurposing of the data. Researchers and analysts can build upon the data producer's work or adapt visualizations for new contexts, extending the impact of the original survey results. This collaborative approach fosters an environment where knowledge and tools are freely exchanged, driving innovation and improving the quality of subsequent studies.

62. Furthermore, the published scripts can serve as valuable training materials for students and early-career researchers. By learning from real-world examples of data analysis and visualization, they gain practical insights into methodological approaches. These scripts provide a hands-on learning experience, demonstrating how theoretical concepts are applied in practice.

63. To ensure that code is replicable, reproducible, and user-friendly for third parties, follow these best practices. Start by documenting the exact versions of input data files used. This step supports reproducibility when third parties have access to the same data, and it clarifies any differences in results when the exact data version is unavailable. Use a version control tool like GitHub to track code updates and modifications, and to manage changes. Specify exact versions of required software or packages, ideally through a configuration file such as requirements.txt in Python or sessionInfo() in R, to prevent compatibility issues.

64. Include in-line comments throughout scripts to explain the purpose of key functions and add a header section for each script that specifies the author(s), date, and purpose of the script. If the codebase includes multiple files, provide clear instructions on the sequence in which scripts should be run. Finally, follow a

[12] Publishing code is not yet a common practice in official statistics. However, it is becoming more common in academia and international organizations. For example, see the Inter-university Consortium for Political and Social Research (ICPSR) OpenICPSR initiative at https://www.openicpsr.org/openicpsr/search/studies, or the World Bank's reproducibility catalogue at https://reproducibility.worldbank.org/index.php/home.

coding style guide, such as PEP 8 for Python or the Tidyverse style guide for R, to enhance readability and make the code easier for others to understand and use.

65. There are several ways to share code, including statistical agencies own websites or online repositories like GitHub, as well as within research papers or accompanying documentation. When sharing code, it is important to provide clear instructions and documentation to help users understand and use the code effectively. Additionally, considering licensing options can help to protect intellectual property while promoting open access and collaboration. Best practice should be followed in the production and documentation of the code (British Ecological Society 2017).[13]

66. One must however note a few limitations regarding the dissemination of programs and scripts:

- **Compatibility with anonymized microdata**: The agency's programs and scripts will have been developed to run on non-anonymized microdata. Running these tools on an anonymized version of the dataset may yield different outcomes or may fail to execute if certain variables have not been included in the disseminated microdata. Therefore, providing clear documentation of these limitations, along with recommendations for adapting the programs and scripts for anonymized datasets, is crucial.

- **Confidentiality of anonymization processes**: The scripts used for anonymizing the microdata should not be shared publicly, as they may contain sensitive information that could potentially be exploited to reverse the anonymization process.

10.3.2. API access: Data as a service

67. Disseminating survey results via an API (Application Programming Interface) adds significant value to traditional data dissemination practices. An API allows users to access and interact with survey data programmatically, turning data dissemination into a flexible and responsive data service. This complements the more traditional data products—such as reports, tables, and visualizations—that statistical agencies produce, by offering a dynamic and customizable means of accessing data. This modern approach caters to the growing demand for flexible and customizable access to data, making survey results more actionable, relevant, and impactful.

[1] **Expanding access and usability**. APIs provide a direct, automated way for users to access survey data efficiently, removing the need to manually download and process large datasets. This is particularly valuable for sophisticated users such as developers, researchers, and analysts who can integrate the data directly into their own tools, applications, and workflows. The ability to retrieve only the specific data they need reduces time spent on manual data handling and allows for more efficient analysis.

[2] **Offering data as a service**. The API model integrates survey data into broader analytical ecosystems and supports the development of innovative tools and applications that extend the utility and reach of the data. While traditional dissemination products like reports and tables are static snapshots of survey results, an API transforms the data into a data service—a live, constantly available source of information that users can query whenever they need updates or specific subsets of data. This shift from static products to dynamic data services aligns with modern expectations for on-demand, customizable data access. It allows users to build applications,

[13] See also the World Bank guidance on reproducibility: https://dimewiki.worldbank.org/Reproducible_Research

dashboards, or analytics platforms that can automatically refresh with the latest available data, improving the utility and longevity of the data.

[3] **Facilitating data integration**. APIs make it easier to integrate survey data with other datasets and systems. Users can combine the survey results with data from other sources (such as economic indicators, geographic data, or administrative records) to conduct more comprehensive analyses. This interoperability enhances the potential for innovative use cases, such as building real-time monitoring systems, visualizing trends over time, or creating predictive models that use survey data as a key input.

[4] **Customization and flexibility**. Disseminating data via API provides greater flexibility for users to access the specific variables or time periods that are most relevant to them. Rather than downloading entire datasets, users can query the API for targeted data extracts, reducing the burden on them to filter, clean, or process large datasets. This customization is particularly valuable for organizations with specific data needs, allowing them to focus on the survey variables most critical to their analysis.

[5] **Supporting innovation**. APIs enable developers, businesses, and researchers to build innovative applications that leverage survey data in new ways. For example, APIs can support the development of mobile apps, data visualization tools, or automated reports that offer real-time insights into the survey results. This creates opportunities for new data products, such as live dashboards or personalized insights, that can extend the reach and impact of the survey data far beyond traditional users. Enabling API access to the survey database is also a key step in making the survey data AI-ready. Providing a standardized, programmatic interface unlocks the data potential for seamless integration with AI tools and platforms.

[6] **Reducing burden on users and agencies**. For both users and statistical agencies, APIs reduce the burden of manual data handling and distribution. Users no longer need to repeatedly download data files, and agencies can automate data dissemination.

68. Making data available via an API requires preparing data storage to ensure efficient access and management. This involves storing the data in a SQL or NoSQL database, depending on the nature of the data and usage requirements. SQL databases, like MySQL or PostgreSQL, are commonly used for structured data with defined relationships, as they support powerful querying and transactional consistency. NoSQL databases, such as MongoDB may offer greater flexibility and scalability when dealing with large volumes or rapidly changing datasets. Once the data is properly stored, an API server and endpoints can be set up to retrieve the data, with endpoints designed to deliver specific subsets or summaries of the data as needed. The database and API integration must be carefully optimized to ensure quick response times and efficient data handling, especially under high traffic. In addition, database indexing, caching, and optimized query structures play a key role in enhancing API performance and ensuring data is delivered accurately and swiftly to users. Security is essential, so authentication mechanisms (such as API keys) should be implemented as relevant to manage access and protect sensitive information.

10.3.3. Scrollytelling

69. Scrollytelling is a powerful method of disseminating survey results that combines storytelling with interactive visualizations to create an engaging, immersive experience. This approach is particularly valuable for users who seek to explore complex datasets in a more intuitive and guided manner.

[1] **Narrative-driven insights**. Scrollytelling transforms raw data into a cohesive narrative, guiding users step-by-step through key findings and insights from the survey. This method allows data to

be presented in a way that emphasizes its context, relevance, and implications, making it easier for users to grasp the main messages. By linking data points to real-world stories and scenarios, scrollytelling helps users understand the broader significance of the survey results.

[2] **Interactive visualizations**. One of the core strengths of scrollytelling is its ability to integrate interactive visualizations into the narrative. Charts, graphs, maps, and infographics are dynamically embedded into the scrolling experience, allowing users to interact with the data in real-time. This interactivity helps users explore different dimensions of the data, drill down into specific segments, and customize their view of the survey results. For sophisticated users, this adds an extra layer of analytical depth.

[3] **Immersive experience**. Scrollytelling creates an immersive experience by seamlessly combining text, visuals, and interactive elements into a single, continuous narrative. As users scroll through the story, visualizations animate, data updates in real-time, and the content evolves in response to user interaction. This immersive approach not only keeps users engaged but also helps them retain more information by actively involving them in the discovery process.

[4] **Bridging data and storytelling**. While traditional survey products like reports and tables focus on presenting the data, scrollytelling bridges the gap between data and storytelling. It turns data dissemination into a more personalized and relatable experience, appealing to users who prefer to learn through narratives. By highlighting key data-driven stories, agencies can emphasize the human or societal impact of the survey results, making the data more accessible and meaningful.

[5] **Customizable user journey**. In a scrollytelling format, users can control their own pace and level of engagement. Sophisticated users who need more detail can explore the data deeply, while others may focus on the overarching story. This customization enhances user satisfaction and ensures that different types of users can derive value from the same product. Agencies can design multiple entry points within the scrollytelling narrative, offering different layers of complexity to suit various audiences.

[6] **Supporting data literacy**. Scrollytelling also promotes data literacy by helping users understand complex datasets through visual storytelling techniques. For users who are less familiar with statistical analysis but are interested in understanding trends and implications, the narrative format helps demystify the data. For more sophisticated users, scrollytelling adds clarity and facilitates the exploration of the underlying data without overwhelming them with technical jargon or static tables.

70. Creating a scrollytelling product to disseminate survey results involves blending storytelling with dynamic data visualizations to engage users as they scroll through insights. This process requires collaboration between subject matter experts and technical professionals. Subject matter expertise is crucial to interpret the survey results accurately, identify the most relevant findings, and frame them in a narrative that resonates with the target audience. The narrative is then supported by data visualizations that make complex findings accessible and visually appealing. Technical expertise is essential to design and implement interactive, responsive elements that guide users through the story in an intuitive flow. This often involves using tools like JavaScript libraries, data visualization platforms, and web development frameworks to bring the story to life. Effective scrollytelling also requires skills in user experience (UX) design to ensure seamless navigation, as well as graphic design and data storytelling to craft visuals that enhance comprehension and engagement.

10.3.4. Interactive tools and dashboards

71. Developing interactive tools and dashboards to disseminate key survey results can significantly enhance user engagement and data exploration. These dynamic platforms allow users to visualize and manipulate data in real time, making complex information more accessible and understandable. By enabling users to filter, compare, and analyse survey findings according to their specific interests and needs, interactive tools foster a deeper understanding of the data and its implications. Additionally, these tools can cater to various user groups, from policymakers to researchers and the general public, thereby increasing the relevance and impact of the survey results.

10.3.5. ChatBots

72. AI tools present exciting opportunities for enhancing user engagement with survey analytical reports and briefs. By developing chatbots, statistical agencies can enable users to interactively converse with survey findings, allowing them to ask questions, seek clarifications, and obtain tailored insights in real time. This conversational approach not only enhances user experience but also democratizes access to complex data. Additionally, AI can facilitate the automatic conversion of analytical outputs into podcasts, making survey findings more accessible to a wider audience. By transforming written reports into audio formats, users can engage with the content while on the go, thereby increasing the reach and impact of the survey results. These innovations can significantly enhance the dissemination of survey findings, making them more user-friendly and widely consumed.

73. Creating a chatbot to disseminate household survey results involves developing an interactive tool that allows users to engage with data in natural language, adapting responses to users' literacy levels for broader accessibility. This approach offers key advantages, as it enables users to ask questions and receive tailored information, making complex survey findings more understandable and relevant. However, there are also inherent risks, as AI-driven chatbots can sometimes misinterpret user questions or provide inaccurate responses due to limitations in natural language processing, especially with nuanced or complex data. To mitigate these risks, thorough testing is essential, ideally incorporating a "red teams" approach where diverse groups of testers challenge the chatbot’s responses and probe for weaknesses in interpretation or misinformation. Including disclaimers in responses is also critical to remind users that, while the chatbot provides useful information, it may not cover all nuances of the data and should not replace expert consultation for more in-depth insights. This combination of thoughtful design, rigorous testing, and clear communication ensures that the chatbot serves as a reliable and user-friendly tool for disseminating survey results.

74. A safe approach for deploying chatbots can be to first deploy it for internal use for the agency’s own internal staff, before any attempts of public facing chatbots.

10.3.6. Podcasts and audio products

75. By transforming written reports into audio formats, users can engage with the content while on the go, thereby increasing the reach and impact of the survey results. Podcasts and other audio products can be a valuable tool for disseminating survey results to a wide audience, especially in countries with low literacy. Audio formats are accessible to people who may not be able to read or write. They can be consumed on a variety of devices, including smartphones and radio receivers. Podcasts and audio products can be used to provide informative and engaging content about survey findings, including key trends, insights, and implications. Podcasts can be produced in a variety of formats, including interviews, lectures, or storytelling. It is essential to ensure high-quality audio and a clear and engaging narrative. When producing podcasts or audio products, it is important to consider the target audience and tailor the content accordingly. This may involve using simple language (including local languages of the communities targeted, when

relevant), avoiding technical jargon, and providing clear and concise explanations. Collaborate with professional podcast producers and radio station journalists to create high-quality content and reach a larger audience. Additionally, it is essential to ensure that the audio quality is high, and that the content is well-structured and engaging.

76. Podcasts can now be automatically generated using AI tools by converting text documents into audio files. The process involves loading a document—for example a survey brief or report on main findings—into an AI system, which then uses advanced text-to-speech (TTS) technology to produce a narrated audio version. Tools like NotebookLM allow users to generate podcasts with minimal effort. At this stage, customization options are limited, and the primary way to control the output is by editing the input text. No technical expertise is required, making the tools accessible to a broad audience. These tools are highly effective, producing clear and engaging audio content, and advancements in the field promise more control and language support in the future.

77. While these tools are efficient, a rigorous quality assurance process is essential before publishing any audio content. The authors of the source document, organizational management, and communication specialists should collaborate to review and approve the generated podcast. This ensures the content of the audio aligns with organizational messaging, maintains accuracy, and resonates with the intended audience.

10.3.7. Videos

78. Producing videos in local languages to share survey results can effectively boost public understanding and engagement. Videos can simplify complex information, making it more accessible to a diverse audience by using visuals, animations, and straightforward language. By adapting content to local languages and cultures, and involving community leaders, statistical agencies can make findings resonate with a broader population. Producing videos in multiple languages promotes greater awareness, informed discussions, and evidence-based policies that meet community needs.

79. Survey results videos can be produced in-house, outsourced, or created using online tools and AI. In-house production allows full control, while outsourcing saves time and resources. Online tools and AI offer cost-effective and professional options with minimal effort. Collaborating with community members can enhance relevance and impact.

80. To ensure high-quality survey results videos, implement a robust quality control system. Use final, verified data, and make videos clear, concise, and accessible to all viewers. Ensure professional production and good audio and video quality. Promote through various channels to reach a wider audience and drive engagement. Videos can be disseminated through various channels, including:

- **TV:** Broadcasting videos on national or regional television channels can help to reach to a wide audience.
- **Web:** Publishing videos on the official website of the producing agency, as well as on social media platforms like YouTube, Facebook, and X. Web videos allow for wider distribution and targeted outreach to specific groups. Creating a dedicated YouTube channel for a survey could make sense – beyond the dissemination of survey results. Featuring a variety of video formats and content, such as explainer videos, interviews, and infographics can help tailor different messages to different audiences.
- **Partnerships with other organizations:** Collaborating with other organizations, such as NGOs, community groups, or educational institutions, helps promote the videos and reach a broader audience.

- **Local events:** Showcasing videos at local events, such as community meetings, development partners workshops, or conferences, to engage directly with the target audience.

10.3.8. Qualifying and understanding data usage

81. In addition to traditional user satisfaction surveys and web analytics, AI tools including large language models (LLMs) can be leveraged to monitor how survey data is cited and utilized in various documents, such as academic papers, reports, and policy briefs. These tools can:

- **Extract information.** AI can analyse documents to identify instances where survey data is referenced, providing valuable insights into the impact and relevance of the data across different fields.
- **Classify use cases.** AI can categorize the contexts in which the data is applied (e.g., policy analysis, economic research), helping organizations understand the broader applications and significance of their survey data.

82. However, a significant challenge in implementing such applications is the inconsistent citation practices among data users. Often, data are under- or inaccurately cited, making it difficult to track the precise usage of specific datasets. To enhance the traceability of disseminated survey results, it is recommended to:

- **Define citation requirements.** Establish clear guidelines outlining how all survey data products should be cited by users. This ensures proper credit is given to the original source and facilitates more effective tracking of data usage.
- **Create Digital Object Identifiers (DOIs).** Assigning a unique DOI to each core data product (e.g., the survey report, or the public use micro-dataset) simplifies the process for users to cite datasets accurately and enables the organization to monitor how the data is referenced across various platforms and publications.

10.4. Resources

UNECE (2018) Guidance on Modernizing Statistical Legislation. Available at https://unece.org/DAM/stats/publications/2018/ECECESSTAT20183.pdf

United Nations (2014) Fundamental Principles of Official Statistics, Available at https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

United Nations (2019) National Quality Assurance Frameworks, Manual for Official Statistics. Available at https://unstats.un.org/UNSDWebsite/data-quality/user-manual

UNECE Making data meaningful, Part 3: a guide to communicating with the media. https://unece.org/DAM/stats/documents/writing/MDM_Part3_English.pdf

UNECE Strategic Communications Framework for Statistical Institutions https://unece.org/sites/default/files/2021-06/ECECESSTAT20211.pdf

Web Content Accessibility Guidelines (WCAG). https://www.w3.org/WAI/standards-guidelines/wcag/

Data Documentation Initiative (DDI) Alliance. DDI-Codebook 2.5. https://ddialliance.org/Specification/DDI-Codebook/2.5/

Olvier Dupriez and Ernie Boyko (2010). Dissemination of Microdata Files - Principles, Procedures and Practices. IHSN Working Paper No 005. Available at http://www.ihsn.org/home/node/120

World Bank (2023). Atlas of Sustainable Development Goals. https://datatopics.worldbank.org/sdgatlas?lang=en

Web Analytics tools:

Google Analytics: https://developers.google.com/analytics

Matomo: https://matomo.org/

British Ecological Society (2017) A Guide to Reproducible Code. Available at https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Reproducible-Code-2019.pdf

World Bank, Reproducible Research website. https://dimewiki.worldbank.org/Reproducible_Research

Google, notebookLM application: https://notebooklm.google/

Part II

of the

Handbook of Surveys on Individuals and Households

Foundations and Emerging Approaches

United Nations
New York

CHAPTER 11
CHILD-CENTRIC SURVEYS

Shane M. Khan (UNICEF)
Attila Hancioglu (UNICEF)
Turgay Unalan (UNICEF)
João Pedro Azevedo (UNICEF)

CHAPTER 11
CHILD-CENTRIC SURVEYS

11.1. Introduction and the need for child-centric surveys

1. Children are a specific, age-defined segment of the population whose needs, as outlined by the UN Convention on the Rights of Children (UN CRC)[1] , include civil, political, economic, social, and cultural rights. Collecting data on such an encompassing set of domains can be challenging, and no single household survey or data source can fulfil all data needs. Multi-topic and thematic surveys have become the preferred tools to generate data on the situation of children given the capacity of surveys to capture a wide range of indicators.

2. However, not all multi-topic surveys capture the full range of child indicators nor do they all follow globally agreed-upon survey methodologies and standards. Hence, classifying household surveys as "child-centric" can increase the precision about how well household surveys deliver the needed indicators for children. Child-centric surveys centre the child as the fundamental unit from which all methodological decisions are made. Child-centric surveys generate a package of data on the child and the child's situation. They must generate data for child-, adolescent- and youth-specific indicators, and reflect the rights cited by the UN CRC and those that are part of the global development agenda, such as the Sustainable Development Goals (SDGs). Child-centric surveys must represent all children in the household population and adhere to global standards for the best and latest methodologies to collecting, analysing and disseminating child data.

3. Meeting all of these criteria may be difficult for a single survey to achieve. Consequently, categorizing a survey as child-centric may rely on some level of subjectivity and should be done using a continuum, with some surveys being more child-centric than others, based on the stated criteria. Consider the USAID-supported Demographic and Health Surveys (DHS) which generates data a wide range of development indicators for young children. The DHS, however, by definition, places more attention on demographic and health data especially for women and men of reproductive ages, and by design, does not collect data on children whose mothers are deceased or living elsewhere (for certain indicators). Child Labour Surveys (CLS), on the other hand, collect data on all children but do so for mainly labour issues. While both surveys have elements of being child-centric, both have deficiencies in representing children or coverage of issues pertinent to the child. Neither survey considers the child as the primordial deciding factor for survey decisions. Currently, the UNICEF-supported Multiple Indicator Cluster Surveys (MICS) meet most child-centric survey criteria. MICS centres children in the design and implementation of surveys, generating child-specific indicators for each phase of a child's life and representing all children in a household population. Nevertheless, a drawback of MICS as a child-centric survey is that certain content such as the health and nutrition of 5–17-year-olds is not collected in MICS, and several indicators on water and sanitation are not child-specific. Instead, these are household-level indicators applied to children.

11.2. Needs, scope, and business case for child-centric surveys

4. Child-centric surveys hold significant appeal, relevance, and importance with national governments and international development partners as they provide key child indicator data for tracking development plans. MICS, an example of a child-centric survey, plays a crucial role in monitoring 33 SDG indicators, the vast majority of which are child-specific[2]. The successful execution of child-centric surveys uniquely relies on the expertise of the national statistical offices (NSOs) complemented by the active involvement of technical specialists from various government ministries to ensure its relevance and use. For example, the execution of MICS surveys draws upon experts from the NSO with support from ministries of education, health, social affairs, other line ministries, UNICEF, and other development partners. In collaboration, these

specialists identify and prioritize the data requirements for the survey, define its scope, including the selection of indicators and the specific age groups of children to be studied. Such a wide group of stakeholders contributing to a singular effort can bring challenges. One of the best practices is to define a clear set of roles and responsibilities and modes of working for partners in the collaboration. This is often formalized in an agreed upon "Terms of Reference" which becomes the formalized governance document for the survey.

11.3. Survey management

5. While ethical review board approvals are common for many surveys, collecting data on children raises unique ethical issues that must be taken into account. The MICS surveys provide a protection protocol document that centres the ethical needs of children in planning and implementation. The protocol allows survey implementers to identify anticipated or actual risks to children, potential harms and benefits to them, and outlines informed consent, privacy and confidentiality concerns. The protocol also identifies measures and procedures to address or mitigate against these risks. Note that ethical review boards may also require additional child safeguarding measures for the survey.

11.4. Questionnaire design

11.4.1. Topics in a child-centric survey

6. Questionnaire content and compliance with international standards are key to defining how child-centric a survey is. Child-centric surveys are required to cover indicators specific to children, adolescents, and youth. Indicators must reflect each phase of a child's development, pointing out the gaps in coverage of key interventions, developmental milestones and other outcome level indicators. **Table 11.1** provides a recommended set of topics for child-centric surveys, disaggregated by age of the child. Many topics in the table are SDGs or related to other global development goals and cover all major sectors of global importance. An ideal set of recommended topics should cover all topics relevant to the child's life course, though for many subject areas, prioritized topics, standardized indicators and methodologies do not exist yet or still need further global agreement.

11.4.2. International standards

7. Another defining feature of child-centric surveys is compliance with international standards. Household survey experts understand that there are numerous approaches to measure the same indicator. However, using standard approaches that are tested, validated, and accepted by the larger international community produces results that are valid, reliable, and comparable across countries.

8. Global questionnaire recommendations on child indicators cover indicator and questionnaire formulation including specifying age groups and the respondent. The child functioning module used by the MICS surveys (developed by UNICEF and the Washington Group Module on Child Functioning), for example, details a set of questions and response categories that were extensively tested and validated across a number of countries. The module includes questions that are specific to children ages 2–4 years and 5–17 years and identifies the respondent as the child's mother or caregiver (if the mother is deceased of living elsewhere). Modifications such as reducing the number of questions or changing the respondent to any adult in the household result in an underestimation of prevalence of child functioning[3].

11.4.3. Questionnaire adaptations for children

9. When children are the respondents to a survey, questionnaire text must be tailored to minimize measurement error at each stage of the survey response process. Survey questions that are age appropriate and easily understood by children aid in comprehension and retrieval stages of the response process, and

thus minimize errors at these stages. The questions should undergo rigorous testing and validation processes and can include cognitive testing, validation of survey responses against "gold standards", interviewer and respondent observations in the field, and expert reviews. Collectively, these techniques can provide a body of evidence on how well survey questions are adapted for children.

Table 11.1. Recommended set of topics for child-centric surveys.

Sector	Household	Mother/birth	0–4 years	5–17years	18–24 years
Health	Clean energy	Antenatal care Delivery care Post-natal care Anaemia	Vaccinations Post-natal care Treatment of diarrhoea Symptoms of ARI		Symptoms of depression and/or anxiety
Nutrition			Anthropometry Exclusive breastfeeding Infant and young child feeding		
Education			Attendance to early childhood education	Foundational learning	
Demographics			Under-five mortality		
Water / Sanitation	Clean water Clean sanitation Water quality				Menstrual hygiene
Poverty / Gender	Social transfers		Health insurance	Children's time use Health insurance	Violence against women Health insurance
Protection	Orphanhood Living arrange-ments	Early child-bearing Birth registration	Child discipline	Child discipline Child labour Child marriage Violence against children	NEET (Not in Employment, Education) or Training
Development			Early childhood development index Child functioning	Child functioning	

10. The MICS surveys interview mainly adults age 18 or older. However, children ages 15–17 and 7–14 are also interviewed. In both cases, adult consent and child asset are sought. Children 7–14 years, for example, are given a reading and numeracy assessment (the Foundational Learning or "FL") module. This module requires a number of deviations from the usual MICS methodology. Apart from interviewer training and field supervision requirements (covered later), the FL module requires that interviewers provide a printed booklet for children to focus on the assessment task. The questions in the FL module were developed and tested for children age 7–14 specifically considering the level of vocabulary and complexity of sentences needed at that age. Each survey that implements the FL module is expected to pre-test the module to ensure comprehension by children. Such deviations are useful considerations and methods that household surveys can use to adapt questionnaires for children.

11.4.4. Pitfalls in applying adult indicators to children

11. Household surveys usually include household and adult indicators, and may provide disaggregates of these indicators applied to children or for children. Such an approach is at times useful but limited as it does not wholly and accurately reflect the breath of child rights and may erroneously assign adult indicators and questionnaire methodologies to children. Multi-dimensional poverty estimates, for example, use a blend of household and individual data. But for child poverty estimates, recommendations state that estimates should be measured for the individual child, and not a disaggregation of a household measure[4]. Consider the problem of equivocally stating that a child has no education because the head of household has never been to school. Child-centric surveys, in contrast, prioritize the child's own well-being and are therefore centred on child-level indicators. In the case of education, for example, the child's own educational level should be collected. Nevertheless, household and adult indicators can provide contextual and environmental data on children and should be included in child-centric surveys, though not at the expense of child indicators.

11.5. Sample design

11.5.1. Sample coverage

12. Coverage is a defining principle of child-centric surveys. Child-centric surveys must sample and represent all children in household populations, from birth to age 24 years, regardless of the geographic focus of the surveys. For instance, a household survey that covers all children in households in an urban area is a child-centric survey, regardless if children themselves or their caregivers are interviewed.

11.5.2. Improving inclusion of all children through oversampling

13. Covering all children in a population can require additional methodological conditions, apart from a complete sample frame. In low-fertility countries, a typical probability sample of households will yield insufficient numbers of children for calculating specific child indicators. This can be remedied by oversampling households with children, a method systematized by the MICS surveys. During the listing phase of the survey, which updates the sample frame in selected clusters, additional data on the ages and dates of birth of residents are recorded. Households with children (specifically, those with children under five years of age) are then sampled at a higher interval than households without children. Results of this approach show that the method significantly increases the number of cases of children in a sample across a number of countries[5].

11.6. Data collection

11.6.1. Consent and assent when interviewing children

14. Child-centric surveys that directly interview children must seek assent from them, in additional to prior explicit informed consent from parents or guardians. These are standard ethical recommendations. In MICS, this is operationalized by using a specific customized assent statements for children. The child assent for ages 7–14 is administered for the Foundational Learning Skills module of the Questionnaire for Children age 5-17. Firstly, adult consent is sought for the Questionnaire for Children age 5–17. If adult consent is given, then, assent is asked for each child selected for interview. An interviewer reads an introduction to the child, asks a number of friendly questions to create trust with the child, and when the child appears comfortable, the interviewer reads the assent statement. After the assent statement is read, the child is able to continue or stop the interview. This statement is tailored to the age of the child to ensure comprehension and emphasizes that voluntary nature of the survey, including the ability of the child to stop the interview without any consequences. Specific consent and assent statements need to be built into questionnaires and programmed into data collection software. Despite the addition of child assent and other changes to MICS questionnaires, response rates have not significantly reduced over time[6]. Household surveys wishing to collect data on children can adapt this methodology of parental/guardian consent and child assent to their surveys.

11.6.2. Implications of ethics for survey implementation

15. During pre-testing and implementation of surveys that directly interview children, surveys must balance children's privacy and confidentiality with protection from abuse. Children should be interviewed in a location that is visible to the parent or guardian while not being overheard. If safety and welfare risks are detected in the survey, survey managers and interviewers have an obligation to act in accordance to the ethical standards for the survey.

11.6.3. Interviewer criteria

16. While household surveys usually have criteria to identify, employ, and train adult interviewers, child-centric surveys require additional training protocols. Interviewers who will work with children need to be carefully screened and selected to ensure that they have experience working with children and do not pose a threat to a child's well-being.

11.6.4. Adjustments for training workers

17. Training also requires additional types of specialized staff to train interviewers to work with children. Child experts from the ministry of health and other government bodies that work with children should provide inputs into the training and include any national guidelines on working with children. Field supervisors and interviewers must be trained to build rapport with children, listen for verbal and non-verbal cues that can indicate distress, and proactively consider the child's well-being. In MICS and other household surveys that collect anthropometric child data, specialized measurers are typically employed to take height and weight measurements of children.

11.7. Weighting

18. Child-centric surveys also need child-centric weights to produce estimates of child-level indicators for different ages. In MICS, for example, weights are adjusted for over-sampling of children under-five (if implemented), and weights corresponding to specific age groups are produced. These include an under-five weight, a 5–17 weight and a weight for children 7–14. Each of these correct for non-response and for variable selection of children of different ages.

11.8. Conclusions

19. Child-centric surveys serve a primary purpose of monitoring the situation of all children, charting their development from birth to adulthood and providing data disaggregates to identify children left behind. Like economic surveys, which are needed for core economic planning, child-centric surveys are necessary tools for social planning. Continued support and adoption of child-centric survey approaches herein discussed through the context of MICS will vastly improve the data landscape for children and propel governmental planning towards achievement of universal child rights.

11.9. References

1 Convention on the Rights of the Child text | UNICEF. https://www.unicef.org/child-rights-convention/convention-text (accessed Dec 13, 2024).

2 Tools. https://mics.unicef.org/tools (accessed Dec 13, 2024).

3 UNICEF. Guidance note on integrating the module on child functioning in demographic and health surveys. New York, NY: UNICEF, 2023 https://data.unicef.org/wp-content/uploads/2023/06/Guidance-note-on-integration-of-CFM-in-DHS-Final-May-2023-1.pdf (accessed Dec 13, 2024).

4 UNICEF. Child poverty profiles: Understanding internationally comparable estimates. UNICEF DATA. 2021; published online Dec 10. https://data.unicef.org/resources/child-poverty-profiles-understanding-internationally-comparable-estimates/ (accessed Dec 13, 2024).

5 Megill D, Khan S, Hancioglu A. Oversampling of children under-five in low fertility settings. New York, NY: UNICEF, 2018 https://mics.unicef.org/sites/mics/files/2024-05/MICS%20Methodological%20Paper%207.pdf.

6 Response rates in household survey programmes: An analysis of MICS and DHS surveys since the 1990s. New York, NY, 2024 https://www.youtube.com/watch?v=KjvYk2GmE9Q (accessed Dec 13, 2024).

This page has been intentionally left blank.

CHAPTER 12
MAINSTREAMING A GENDER PERSPECTIVE

Jessamyn Encarnacion (UN Women)
Hemalata Ramya Emandi (UN Women)
Papa Seck (UN Women)
Aurelie Acoca (UN Women)
Maria Isabel Cobos Hernandez (United Nations Statistics Division)
Iliana Vaca Trigo (United Nations Statistics Division)

CHAPTER 12
MAINSTREAMING A GENDER PERSPECTIVE

12.1. Introduction

1. As gender issues gain increasing prominence in national and global commitments and agendas, the demand for gender statistics[1] and integrating a gender perspective into statistical systems continues to grow. These are distinct yet interconnected concepts. **Gender statistics** go beyond sex-disaggregated data; gender statistics are designed to capture and reflect the differences and inequalities between women and men across all dimensions of life.[2] Gender statistics provide crucial evidence for research, policy development, and programme implementation while improving statistical systems to reflect better the diverse activities and characteristics of women and men.[3]

2. **Mainstreaming a gender perspective in statistics** involves systematically addressing gender issues and mitigating gender-based biases throughout the production and use of official statistics, and across all stages of the data value chain[4]. This overarching principle must be applied to all surveys – not only those focused specifically on women or gender issues but to all types of surveys and data collection efforts. For instance, ensuring gender balance in survey design teams can influence the framing of questions, improving their inclusivity and relevance for all genders, and some surveys dealing with sensitive topics, such as violence against women and girls, require all-female survey team to ensure data quality.[5] Similarly, collecting data disaggregated by sex and other intersectional factors, such as income, age, disability, or geographic location, enables a deeper understanding of gender-differentiated disparities and inequalities with an intersectional lens. Mainstreaming a gender perspective is thus deliberately emphasized and thoroughly addressed in this Handbook, particularly in this Chapter. This Chapter provides guidance to challenge biases inherent in general survey processes[6] and traditional survey practices, addressing these societal realities and avoiding perpetuating such biases.

3. This Chapter illustrates how to mainstream a gender perspective into surveys throughout the survey process, drawing from lessons learned and best practices of countries and the international statistical community. The Chapter consists of the following sections, aligned to the core chapters of the Handbook (Chapters 2-10):

---

[1] For example, the transition from the Millennium Development Goals (MDGs) to the Sustainable Development Goals marked a significant increase in the emphasis on gender data. Under the MDGs, gender equality was primarily addressed through Goal 3, with four indicators. In contrast, the SDGs adopted a more comprehensive framework with 52 gender-specific indicators spread across Goal 5, which is dedicated entirely to achieving gender equality and women's empowerment, as well as in other 10 of the 17 Goals.

[2] UN Statistics Division. 2016. Integrating a Gender Perspective into Statistics.

[3] UNECE and World Bank Institute. (2010). Developing Gender Statistics: A Practical Tool

[4] Data2X and Open Data Watch. The Data Value Chain: Moving from Production to Impact. https://opendatawatch.com/reference/the-data-value-chain-executive-summary/

[5] UN Statistics Division. 2014. Guidelines for Producing Statistics on Violence Against Women.

[6] In most cases, these processes follow the Generic Statistical Business Process Model (https://unece.org/statistics/modernstats/gsbpm)

- **Needs, scope, and business case:** Section 12.2 highlights the importance of integrating a gender perspective from the outset, shaping the planning, design, implementation, and evaluation of the survey process.
- **Survey management:** Section 12.3 underscores the need for gender balance and representation in survey project management.
- **Questionnaire design:** Section 12.4 recognizes the need to address gender biases in questionnaire design and framing.
- **Sampling:** Section 12.5 focuses on mitigating gender biases that may arise from under-coverage or over-coverage of target populations caused by deficiencies and biases in the sampling frame or method, which can result in skewed outcomes.
- **Data collection:** Section 12.6 emphasizes using effective data collection methodologies to ensure gender-sensitive and inclusive data.
- **Data processing:** Section 12.7 shows how to ensure that the integrity of the gender data is preserved in the handling, cleaning, and processing to reflect gender-specific insights from survey respondents accurately.
- **Weighting:** Section 12.8 highlights the importance of integrating a gender perspective in the weighting process to ensure that the results accurately reflect the diverse experiences and contributions of women, men, girls, and boys.
- **Analysis:** Section 12.9 underscores the significance of analysing survey data to uncover disparities and inequalities and to comprehensively understand gender dynamics within the surveyed population.
- **Dissemination:** Section 12.10 highlights the importance of communicating survey findings that promote inclusivity and do not reinforce gender stereotypes and biases.

12.2. Specifying needs, scope, and business case

4. The "Needs, Scope, and Business Case" stage (see Chapter 2) starts when a need for a new survey is identified or feedback on an existing or regular survey arises. This stage involves working with stakeholders to understand what specific information is needed now or in the future, suggesting possible data solutions, and creating a survey plan to address those needs.[7] To ensure that relevant gender issues are integrated into the survey objectives, it is essential to incorporate a gender perspective from the outset in the following ways, including as required under the Istanbul Convention[8]:

[7] UNECE. Generic Statistical Business Process Model (GSBPM) accessed at https://unece.org/sites/default/files/2023-11/GSBPM%20v5_1.pdf

[8] The Istanbul Convention, adopted by the Council of Europe Committee of Ministers in 2011, is the first international treaty to define gender as a socially constructed category, shaped by roles, behaviours, and attributes assigned to women and men. It highlights the link between gender equality and eradicating violence against women, recognizing such violence as rooted in historically unequal power dynamics between genders.

12.2.1. Specifying gender-relevant needs, survey objectives, and scope

5. During this stage, NSOs and other data producers within the NSS should consider the following best practices to ensure gender is considered.

6. **Consult gender experts, producers, and users from the outset and throughout the survey process to ensure alignment with national policies and priorities – that is, intended policy action driving gender data production in surveys.** During the needs assessment process (see Chapter 2), gender experts, producers, and users must be involved from the start to ensure surveys address policy needs and can drive evidence-based action. Experts and stakeholders to be consulted may include gender data focal points and policy specialists from national women's machinery[9], ministries, departments, and agencies (MDAs); representatives of inter-agency committees on gender statistics (IAC-GS) (see **Box 12.1**); civil society organizations (CSO); women's advocacy groups; and researchers.

7. The consultation process should explicitly link data collection to national policy goals, ensuring the data not only supports monitoring efforts but also provides the evidence needed to inform and catalyze policy action. Key steps in consultation include stakeholder mapping, effective communication, and collaboration agreements to align data efforts with broader policy frameworks.[10]

8. **Set clear gender-sensitive survey objectives.** Define intentional objectives addressing women's and men's needs, covering areas like unpaid care work, women's participation and decision-making, gender-based violence (GBV), gender and the environment and digital transformation among other topics.[11]

Box 12.1. Mainstreaming gender perspective in national surveys in Mexico through the Specialized Technical Committee on Information with a Gender Perspective

An example of mainstreaming a gender perspective is Mexico, where the National Institute of Statistics (INEGI) and the National Institute for Women (INMUJERES) signed a cooperation agreement in 2001 to produce gender-focused statistics, building on work initiated in 1997. In 2010, the Specialized Technical Committee on Information with a Gender Perspective (CTEIPG) was created to promote gender-sensitive statistics for national policy planning and monitoring. In 2021, another agreement was signed to strengthen the mainstreaming of the gender perspective, focusing on generating, standardizing, and exchanging statistical data through national surveys, censuses, and administrative records. This collaboration supports the National Program for Equality (PROIGUALDAD) and fosters the creation of methodologies, standards, and joint dissemination efforts through conferences and seminars.

Source: ECLAC. (2022). Breaking the statistical silence to achieve gender equality by 2030.

[9] Women's machinery, or national machinery for the advancement of women, is a government unit that coordinates policies to support gender equality and women's empowerment

[10] UN Women and UN SIAP. (2020). Module 4, "User-Producer Dialogue", Gender Statistics Training Curriculum

[11] See the Beijing Declaration and Platform for Action for the list of the 12 critical areas of concerns agreed by governments in 1995.

9. **Identify gender-specific indicators the survey could support for monitoring progress in national and international agendas**. Include indicators aligned with:

- *National priorities*: National priority gender equality indicators (NPGEIs) developed by National Statistical Systems (NSS)
- *Regional frameworks*: Minimum/core sets of gender indicators[12] (e.g., for Latin America and the Caribbean[13], Africa[14], Arab States[15])
- *Global commitments*: Minimum set of gender indicators[16], Gender-specific indicators of the Sustainable Development Goals[17] in the Agenda 2030[18], Beijing Declaration and Platform for Action (BPfA)[19] and the Convention on the Elimination of All Forms of Discrimination against Women (UN CEDAW)[20].

10. To further support in identifying gender-specific indicators across sectors, refer to resources such as the UN Handbook on Integrating Gender Perspective into Statistics, Voluntary National Reviews, National Strategies for the Development of Statistics (NSDS), UNECE Tool for Developing Gender Statistics and other sector-specific resources (e.g., time-use and unpaid care and domestic work, violence against women, environment) (see **Box 12.2**).

[12] The Minimum Set of Gender Indicators endorsed by the United Nations Statistical Commission (decision 42/102) in 2013, and revised in 2021 serve as a guide for national production and international compilation of gender statistics.

[13] The Gender Equality Observatory publishes a set of relevant indicators to monitor progress towards gender equality and the economic, physical and decision-making autonomy of women and girls in Latin America and the Caribbean.

[14] African Development Bank Group, UN Women, and United Nations Economic Commission for Africa. Minimum Set of Gender Indicators for Africa: Phase IV Report. (https://data.unwomen.org/publications/minimum-set-gender-indicators-africa-phase-iv-report)

[15] UN Economic and Social Commission for Western Asia. 2023. Handbook on the Arab Gender Indicator Framework 2023. (https://www.unescwa.org/sites/default/files/pubs/pdf/handbook-arab-gender-indicator-framework-2023-english_1.pdf)

[16] United Nations Statistics Division. Minimum Set of Gender Indicators. https://gender-data-hub-2-undesa.hub.arcgis.com/

[17] UN Women and UN Statistics Division. 2024. The Gender Snapshot 2024. https://www.unwomen.org/en/resources/gender-snapshot

[18] United Nations. The Sustainable Development Agenda. https://www.un.org/sustainabledevelopment/development-agenda/

[19] United Nations. 1995. Beijing Declaration and Platform for Action and the Beijing + 5 Political Declaration and Outcome (Reprinted by UN Women in 2014: https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/CSW/PFA_E_Final_WEB.pdf)

[20] United Nations. 1979. The Convention on the Elimination of All Forms of Discrimination Against Women. (https://wrd.unwomen.org/practice/resources/convention-elimination-all-forms-discrimination-against-women-cedaw)

Box 12.2. Gender mainstreaming in the data collection strategy on entrepreneurship and financial inclusion in Cameroon

Cameroon has made significant strides in integrating a gender perspective into its statistical processes. One key achievement has been the creation of a Permanent Working Group on Gender Statistics within the National Institute of Statistics (NIS). This group acts as a think tank, coordinating gender-sensitive activities across the country's decentralized statistical system, which includes various structures involved in producing official statistics.

Cameroon's approach to mainstreaming gender into household surveys addresses gender disparities through data disaggregation, particularly in areas like women's entrepreneurship, health, financial inclusion, and poverty. Identifying a minimum set of gender indicators has been pivotal in guiding data collection efforts to ensure that gender issues are systematically captured and used to inform policy.

Gender mainstreaming in household surveys directly benefits women by highlighting gender-specific challenges, promoting data-driven policymaking, and ensuring gender equality in economic and social sectors. For instance, the women's entrepreneurship data has led to targeted policies supporting female business owners. By investing in such surveys, Cameroon not only addresses critical gender issues but also enhances the relevance of its data for informed policymaking that benefits women across various sectors.

Source: UN Women. (2019). Counted and Visible: Toolkit to Better Utilize Existing Data to Bridge Gender Data Gaps. https://data.unwomen.org/resources/counted-and-visible-toolkit

12.2.2. Gender-sensitivity of a survey business case

11. Every business case for a proposed survey should clearly articulate *how and why* a gender perspective is mainstreamed into all stages of the survey process, including data collection, processing, analysis, and dissemination. The business case should demonstrate the unique value of the survey by addressing critical gender issues, highlighting the survey's potential to generate actionable insights that benefit women and other underrepresented groups. It is particularly important for example to ensure that the survey's objectives and output are aligned with and can help inform the development, implementation and monitoring of national priorities, including gender equality policies. UN Women's gender data programme, Women Count, presents a business case of surveys promoting stakeholder collaboration and consultations and by highlighting the intended data use of survey results. For example, the 2021 Kenya Time Use Survey was produced through a collaboration between UN Women and Kenya's National Bureau of Statistics (KNBS), in partnership with the State Department for Gender and Affirmative Action, the World Bank, and Oxfam, and in consultation with National Gender and Equality Commission, the Council of Governors, and the University of Nairobi. The data is used to inform the National Care Needs Assessment and to build Kenya's first ever National Care Policy[21].

12.3. Survey management

12. The quality of survey data is intrinsically linked to how the survey is managed. Therefore, integrating a gender perspective in surveys requires management practices that uphold the principles of gender equality

[21] UN Women. 2024. Summary Brief: Kenya's Time-Use Survey and Care Assessment. (https://data.unwomen.org/publications/summary-brief-kenyas-time-use-survey-and-care-assessment)

and inclusivity. When handled in a gender-sensitive manner, surveys generate not only reliable and representative gender data but also address potential biases, optimize resources, and build stakeholder trust. It is not simply about achieving quantitative goals but about changing the institutional and systemic culture so that the National Statistical System, in particular, and the country, in general, can fully deliver on its commitments to gender equality.

13. Following are some recommendations to ensure effective and gender-sensitive survey management:

12.3.1. Equal representation in survey management

14. It is often a case where high-level officials in leadership positions are predominantly men with statistical backgrounds.[22] According to a PARIS21 report[23], in 2021, less than a quarter of top leadership positions (Director General, Chief Statistician) of NSOs were women. Gender parity should be observed in the leadership and membership of the teams to be established (e.g., Steering Committee, Advisory Group, Survey Teams). For example, the United Nations-wide Strategy on Gender Parity sets targets and monitors the following areas: leadership and accountability, senior management, recruitment and retention, creating an enabling environment, and Mission settings. The same principle should be applied in survey management to foster a gender-neutral and inclusive workplace. (See UN Women training resources on *Forging Pathways to Gender Equality in Statistical Leadership*[24] for reference.)

12.3.2. Inclusion of gender experts in teams

15. It is important for gender experts not just to be consulted but meaningfully engaged at every stage of survey planning, including questionnaire development, supervisor and enumerator selection, training manual sensitization, and data analysis. Aside from bringing specialized knowledge and insights, by involving gender experts, survey teams can also strengthen their capacity to produce high-quality data that informs evidence-based policies and promotes gender equality effectively.

12.3.3. Gender-responsive survey budgeting, financing, and resourcing

16. Many people are familiar with gender-responsive budgeting (GRB), but the same cannot be said of gender-responsive statistical—or, in this case, survey budgeting. The latter pertains specifically to the *intentional* allocation of survey resources to produce, analyse, and disseminate gender statistics, where, in some instances, additional data collection and processing may be required. Consequently, training sessions and workshops to support these activities are essential, along with allocating adequate resources to ensure effective implementation.

12.3.4. Stakeholder engagement

17. Establishing structures and mechanisms for stakeholder engagement throughout the survey process is essential to ensure the consistent integration of gender perspectives. These mechanisms facilitate collaboration, foster inclusivity, and enable diverse stakeholders—including gender experts, community

[22] UNFPA. (2013). Genderizing the Census: Strategic approaches to capturing the gender realities of a population

[23] Suchodolska, L. and A. Misra. 2021. The leadership gap in statistics—where are the women. https://www.paris21.org/news-center/news/leadership-gap-statistics-where-are-women

[24] UN Women and Guerrero, Margarita. 2024. Training Courses on Forging Pathways for Gender Equality in Statistical Leadership. https://data.unwomen.org/resources/gender-statistics-training-curriculum

representatives, and policymakers—to contribute their insights at every stage, from survey design to data dissemination.

12.3.5. Approvals

18. The approval process should ensure compliance with existing gender-sensitive laws, policies, and frameworks relevant to the statistical process and survey thematic topics (e.g., violence against women). This alignment reinforces the commitment to upholding gender equality, women's and girls' rights and empowerment principles. It ensures that the survey adheres to ethical and legal standards (e.g., ensuring ethical and safety protocols are followed when conducting sensitive surveys such as violence against women and girls (VAWG) survey).

12.3.6. Gender-sensitive structure and staffing

19. Gender balance should also be aimed at the survey staffing structure. Further, gender sensitization of personnel involved in the survey (e.g., trainers, supervisors, enumerators, data coders, and data editors) is necessary.

12.3.7. Quality assurance

20. Gender sensitivity of the system, process, task, or output is a quality aspect that should be observed, that is, beyond the aspects outlined in the UN National Quality Assurance Framework. UN Women asserts that data—in this case, survey data—is only of quality if the gender perspective is mainstreamed throughout the process.[25]

12.4. Questionnaire design and development

21. Evidence has shown that it is important to collecting data from the appropriate household member(s) and ensuring that gender roles and knowledge are fully considered in survey design. A case in point[26] is maternal and child health surveys, where inconsistencies in responses about childhood illnesses were found when data was collected from different household members, i.e., when men were asked about healthcare expenditures while women were asked about treatment practices. Generally, women provided more comprehensive accounts of childhood illnesses, whereas men tended to recall only more severe cases involving expenditure.

12.4.1. Clarity on concepts and variables

22. To improve data quality and reliability, particularly in low- and middle-income countries, it is essential to recognize and address gender biases in question framing and respondent selection. Below are some recommendations:

23. Existing concepts, definitions, and survey variables should be evaluated to produce unbiased gender-relevant information without being influenced by stereotypes, assumptions, or systemic biases. For example, the designation of household heads frequently assumes that men are the default, overlooking women's roles in decision-making, resource management, and caregiving. Surveys teams should not make assumptions about roles and decision-making dynamics in the household, and where necessary, should instead allow for multiple members to be selected and interviewed.

[25] 2021, UN Women, Counted and Visible Toolkit

[26] Alba, Sandra & Wong, Franz & Bråten, Yngve. 2019. Gender Matters in Household Surveys. Significance. 16. 38-41. Accessed at https://www.researchgate.net/publication/337431301_Gender_matters_in_household_surveys

24. Confusion between “sex” and “gender” persists among producers and users of statistics[27]. The word “sex” refers to biological differences between women and men. Biological differences are fixed and unchangeable and do not vary across cultures or over time. “Gender” refers to socially constructed differences in the attributes and opportunities associated with being female or male and social interactions and relationships between women and men. If the questionnaire is designed to measure them separately, ensure a clear distinction is made between them. Questions on sex, often asked for demographic purposes, maybe “*What is your sex?*” or “*What is your sex assigned at birth?*”. Gender questions, on the other hand, may be stated as: “*What is your gender?*” or “*What is your current gender identity?*” Gender can include categories like male, female, non-binary, or other identities.

25. The distinction between "work" and "employment" often poses challenges because these concepts capture different dimensions of economic activity that are not always adequately addressed or valued. "Work" encompasses a broad range of activities, including unpaid care work, subsistence activities, and voluntary labour, which are often disproportionately performed by women and may be overlooked in traditional surveys focused solely on "employment." Employment, as commonly defined, refers to activities that are remunerated or contribute to formal economic output, thus excluding many informal or unpaid forms of work. This conceptual gap can lead to underrepresentation of women's contributions to the economy and hinder the development of policies aimed at addressing gender disparities.

12.4.2. Ethical and safety protocols must be followed

26. The following should be incorporated in the questionnaire when asking about gender-related sensitive topics (e.g., violence against women and girls, sexual and reproductive health, sexual orientation and gender identity, and asset ownership)[28]: (See Section 12.6.1 on *Safety and ethical considerations being of utmost importance in data collection* for respondents and interviewers.)

- Categorical requests for consent should be included. In surveys done remotely, ask a screening question to ensure the respondent is alone and in a private space. [29]
- Make clear in the consent that the survey is voluntary and that the respondent can stop it at any time.
- When asked sensitive questions, ask whether the respondent feels safe to answer these types of questions.
- The questionnaire should reflect a safety word agreed upon before the interview for the respondent to inform the interviewer in secrecy if it is not safe to respond to survey questions (e.g., when the female respondent in a *remote/telephone-based* VAW survey is no longer alone).
- Use simple response categories that are not indicative of the topic of the survey itself and where a participant can respond with neutral answers such as “yes”, “no”, “agree”, “disagree”, and other

[27] UN Statistics Division. 2016. Integrating a Gender Perspective into Statistics.

[28] UN Women. 2022. Measuring the Shadow Pandemic: Collecting Violence against Women Data through Telephone Interviews. An Evidence-Based Technical Guidance. (https://data.unwomen.org/sites/default/files/documents/Publications/Guidance_VAW_RGA-EN.pdf)

[29] See example on informed consent from the questionnaire on Rapid Gender Assessment on the Impact of COVID-19 on Women’s Wellbeing accessed at https://data.unwomen.org/sites/default/files/documents/Publications/Annex-Phase-II-questionnaire_RGAs-VAW.pdf

similar phrases (rather than requiring the respondent to name the sensitive behaviour or experience verbally). For example, ask a woman respondent whether, "*In the past 30 days, did you ever feel unsafe in your home?*" answerable by Yes or No (or refuse to answer) instead of asking how she feels staying at home.

- Include prompts in the questionnaire to periodically check in with the respondent to ascertain comfort and privacy.

- There should be a script/text on the questionnaire that: i) reassures the respondent that the information provided will be kept confidential; and) offers information about available services and sources of support (local agencies, if possible, i.e., helplines, women's centers, psychosocial support) regardless of incident disclosure. This follows guidelines developed by UN Women and WHO on remote VAW data collection[30], WHO on research on domestic violence[31], and UNSD on VAW surveys[32].

12.4.3. Use of gender-sensitive language and terminology

27. Employing gender-sensitive and -inclusive language helps avoid reinforcing stereotypes or biases and ensures that diverse respondents understand and accept survey questions and materials. To achieve this, it is critical to provide adequate training and sensitization for all survey staff, equipping them with the knowledge and gender sensitivity required to address these issues effectively. The following should thus be considered:

- Using concepts and definitions that adequately reflect gender issues (e.g., unpaid care and domestic work)

- Clarity and wording of questions so that no room is left for interpretation based on the respondents' and (in the case of one-to-one interviews) interviewers' own personal beliefs and social norms

- Avoid use of derogatory language to identify categories of persons (e.g., use 'domestic workers', not 'domestic servants')

- Categories of marital status should be detailed enough to capture various types of informal unions but also mindful of social and cultural norms (e.g., cohabiting/live-in partner)

- Do not use "non-working" when referring to "unemployed" to avoid perpetuating responses or misconceptions about women not working when all forms of work – including unpaid care and

[30] UN Women and WHO. 2020. Violence Against Women and Girls Data Collection during COVID-19. https://www.unwomen.org/en/digital-library/publications/2020/04/issue-brief-violence-against-women-and-girls-data-collection-during-covid-19

[31] World Health Organization (WHO). 2001. Putting Women First. https://www.who.int/publications/i/item/WHO-FCH-GWH-01.1
and WHO. 2016. Ethical and safety recommendations for intervention research on violence against women. https://www.who.int/publications/i/item/9789241510189

[32] UN Statistics Division. 2014. Guidelines for Producing Statistics on Violence against Women – Statistical Surveys. https://unstats.un.org/unsd/gender/docs/guidelines_statistics_vaw.pdf

domestic work often performed by women – are recognized per Resolutions and Guidelines adopted in the 19th International Conference of Labour Statisticians[33].

12.4.4. Use of indirect questions on sensitive topics, especially in remote data collection

28. Indirect questioning techniques can offer a practical solution when asking about sensitive topics and address issues such as underreporting, social desirability bias, and safety concerns. It often creates a more comfortable environment for respondents to share sensitive information. For example, in UN Women's rapid gender assessment on the impact of COVID-19 on women's well-being[34], UN Women utilized specific techniques to collect data on safety in private and public spaces: vignettes and list randomization[35] as indirect questions on respondents' safety in private and public spaces. See also Chapter 15 for other applications of these techniques for measuring sensitive topics related to governance.

12.5. Frame and sampling

29. Sampling strategies must account for diverse household structures and roles to capture women's experiences in various contexts.

12.5.1. Account for intersectional dimensions of gender inequality through targeted sampling

30. **Gender biases can arise from under-coverage or over-coverage of target populations in the sampling frame.** These coverage issues can lead to biased results if the under-covered or overrepresented groups differ systematically in the outcome being measured. For example, because women are less likely than men to be literate, own mobile phones, or have internet access in many developing countries, phone- or web-based surveys that do not adjust for these factors can lead to significant biases. Similarly, in a country where women are more likely than men to be in informal employment, using business registries or employee lists as sampling frames will greatly underestimate the contribution of women to the economy through their informal employment. To overcome the challenge, a supplementary frame would be needed to capture women's and men's informal employment.

31. **Sample sizes are sometimes too small for intersectional analysis.** It is imperative to go beyond simple sex disaggregation to account for intersectional dimensions of gender inequality, such as age, disability, or socio-economic status, aligning with the Leave No One Behind (LNOB) principle. For example, older women with disabilities may face compounded barriers to accessing support services when they experience violence. For a more inclusive and intersectional data collection, oversampling marginalized groups of women or expanding sampling frames to include hard-to-reach populations often excluded (e.g., Indigenous women) may be considered.

32. There will inevitably be trade-offs, including those related to budget, coverage, and other constraints. Prioritization will depend on a careful balance of survey objectives, users' needs, and the intended use of survey results for specific policies and programmes. For NSSs aiming to mainstream gender, it is essential

[33] International Labour Organization. Forms of work: An overview of the new statistical standards. https://ilostat.ilo.org/methods/concepts-and-definitions/forms-of-work/

[34] UN Women. 2021. Rapid Gender Assessment on the Impact of COVID-19 on Women's Well-Being. https://data.unwomen.org/publications/vaw-rga#_dashboardFilterRGA

[35] UN Women. 2023. Innovative Survey Techniques in Collecting Data on Violence against Women During COVID-19: The Example of List Randomization.

to explicitly address these trade-offs and ensure that adequate funding and resources are allocated to achieve this goal effectively.

12.5.2. Individuals as the unit of analysis

33. The unit of analysis determines what data is collected and how a gender perspective is incorporated. Biases can emerge depending on whether the unit of analysis is at the individual or household level. Using households as the unit of analysis can mask gender disparities in access to and control over resources. For example, household income and asset data often fail to capture critical aspects of gendered differences, such as access to resources and intra-household dynamics. To address these distinct challenges, adopting individuals as the unit of measurement can better reflect gender-differentiated access and control of resources. Additionally, to effectively capture intra-household dynamics, it is essential to interview at least two household members of different sexes separately. For example, we can compare whether the husband and wife agree with contraceptive use; or with the ownership of assets[36] (See **Box 12.3**).

Box 12.3. Conceptual framework for measuring asset ownership from a gender perspective: The Evidence and Data for Gender Equality Project by UNSD and UN Women

The UNSD and UN Women's Evidence and Data for Gender Equality (EDGE) project on measuring asset ownership emphasizes the importance of collecting data at the individual level rather than solely at the household level. When collecting individual-level data on asset ownership, a range of statistics can be generated to answer key policy questions relating to three objectives: measuring the gender asset gap, measuring the gender wealth gap, and (in households where more than one member is interviewed), understanding how asset ownership and wealth are distributed among couples or by gender within households.

To ensure data accuracy, EDGE recommends self-reported information on asset ownership, highlighting that proxy responses can lead to significant discrepancies. These approaches uncover gender disparities in asset ownership, informing policies to foster women's empowerment and reduce poverty.

Source: Measuring Asset Ownership with a Gender Perspective. https://unstats.un.org/edge/methodology/asset/

12.5.3. Self-responses

34. Direct responses from individuals enable the collection of nuanced and sensitive information, such as decision-making roles, access to resources, or experiences of discrimination and violence, which may be misreported or overlooked when using proxies. Prioritizing self-responses requires careful consideration of several factors, such as whether it may necessitate additional visits to reach respondents directly, whether clear guidelines must be established when proxies are unavoidable, the training of enumerators in handling sensitive topics, and strategies to accommodate respondents' schedule and availability.

[36] United Nations. 2019. Guidelines for Producing Statistics on Asset Ownership from a Gender Perspective. https://unstats.un.org/edge/publications/docs/Guidelines_final.pdf

12.6. Gender-sensitive survey data collection

12.6.1. Safety and ethical considerations are of utmost importance

35. Protecting the privacy and security of individuals is a fundamental ethical responsibility in data collection, especially when dealing with sensitive information. While smaller groups, such as women in domestic violence situations or LGBTIQ+ individuals in repressive societies, face heightened risks, privacy and safety considerations extend to all populations. Ethical and safety protocols should be in place to ensure respondents' safety and confidentiality, such as ensuring women are alone when responding, allowing the use of a safe word during the survey, checking that speakerphone or call recording is not in use, and providing support resources to all respondents.[37]

36. Experience has shown that, despite the sensitivity of particular surveys, such as violence against women, interviewers can collect reliable and valid data if they are well-trained and empathetic[38]. It is equally important to include gender experts within the training teams to ensure a nuanced understanding of the subject. Survey managers must also anticipate and address potential emotional trauma among interviewers who may be affected by hearing repeated accounts of violent victimization.[39] Protocols should guide managing emotional distress, such as taking breaks or contacting a supervisor or counsellor during fieldwork, to promote ethical and effective data collection. Following the WHO ethical and safety guidelines is essential to safeguarding respondents *and* interviewers.

12.6.2. Ensure comprehensive data collection

37. Comprehensive data collection involves designing surveys that capture diverse population groups' experiences, activities, and realities, ensuring that no significant aspects are overlooked. This approach requires addressing potential biases in data collection, ensuring inclusivity, and incorporating methodologies that reflect the complexities of daily life, particularly in gender-sensitive contexts.

38. An example of this is the collection of time-use data, which is critical for understanding various aspects of women's empowerment.[40] To avoid gender biases in time-use data, UNSD recommends classifying activities for time-use, recording simultaneous activities to reflect the realities of multitasking, particularly for women, and ensuring contextual information—such as whether activities are paid or unpaid and coverage of relevant individual and household characteristics[41].

[37] UN Women. (2021). Measuring the pandemic: Violence Against Women during COVID-19

[38] UN Women. 2022. Rapid Gender Assessment on the Impact of COVID-19 on Violence against Women: Survey Technical Report April – September 2021.

[39] UNECE and World Bank Institute. (2010). Developing Gender Statistics: A Practical Tool

[40] Time allocation patterns; unpaid work; participation in all forms of work; working time, work locations and the scheduling of economic activities; work-family balance; the investment of time in education and health; welfare and quality of life; and intrahousehold inequality.

[41] UNSD. 2024. „Guide to Producing Statistics on Time Use". https://unstats.un.org/UNSDWebsite/statcom/session_55/documents/BG-3k-Guide_to_Producing_Statistics_on_Time_Use-55UNSC_background-E.pdf

12.6.3. Conduct gender-sensitive trainings

39. Due to social norms and power dynamics, enumerators' gender and age can significantly influence data quality, particularly when collecting sensitive information. Gender-sensitive training is essential to equip enumerators with the skills to handle sensitive topics ethically and effectively. This includes understanding local gender issues, cultural norms, and how responses may differ between women and men.

40. For instance, in time use surveys, women may underreport unpaid work, while men may overstate household contributions; women may hesitate to discuss issues of bodily integrity with male enumerators and vice versa. Gender matching—pairing interviewers and respondents of the same gender—can enhance openness and data quality. Training should also address how enumerators' values and beliefs might influence responses, ensuring a bias-free approach to collecting data on gender inequalities.

12.6.4. Leveraging innovative techniques

41. Emerging technologies can complement traditional methods of collecting gender-sensitive data by integrating subjective response data with objective measures through sensors and other electronic devices. Wearable devices such as smartwatches can supplement time-use data, and GPS-enabled survey support can map women's access to resources such as land. This provides cost-effective ways to improve data quality by improving accuracy and reducing subjective response biases.[42] However, these technologies must be applied thoughtfully to avoid reinforcing existing biases and ensure inclusivity. For example, technologies should take into consideration marginalized groups of women faced with limited digital literacy compounded by already prevailing inequality on the digital gender divide[43] (that is, the disparity between women's and men's access to, use of, and benefits from digital technologies).

12.7. Data processing

42. To ensure data neutrality and reflect gender-specific contexts, NSOs and other data producers should avoid data editing practices that can introduce or reinforce gender biases. Examples of what *not* to do include: data imputation based on gender stereotypes (e.g., assuming women are less likely to be employed in certain sectors such as STEM fields or that men are not involved in caregiving roles); bias in outlier treatment by automatically treating gender-atypical responses as outliers (e.g., men reporting high hours of unpaid care work or women in leadership positions) and excluding them from analyses; enforcing binary categories (of women and men) by editing responses outside the binary categories; generalizing household roles by assuming household heads are men and women are caregivers; prescriptive data edits such as in perceiving some of the time-use activities reported as non-productive (e.g., care work), which diminishes women's contributions.

43. By explicitly avoiding these editing practices, NSOs can ensure that data processing respects gender-specific contexts and truly reflects the realities of all genders.

[42] 2019, ISWGHS, at 50th UN Statistical Commission https://unstats.un.org/unsd/statcom/50th-session/documents/BG-Item4c-ISWGHS-E.pdf

[43] UN Women. 2024. Placing Gender Equality at the Heart of the Global Digital Compact. Taking Forward the Recommendations of the 67th Session of the Commission on the Status of Women. https://www.unwomen.org/sites/default/files/2024-03/placing-gender-equality-at-the-heart-of-the-global-digital-compact-en.pdf

12.8. Weighting

44. Calibrating survey weights using gender, alongside other variables such as age, region, and socio-economic status, ensures that final weighted estimates align with known population distributions. This approach prevents the over- or underrepresentation of women or men in the data. For surveys with highly gender-specific content, such as intimate partner violence, creating separate weights for women and men can enhance the accuracy of gender-sensitive analyses. For example, India's National Family Health Survey 5 (NFHS - 5) employed gender-specific weights to address differing response patterns and ensure robust, gender-disaggregated insights[44].

12.9. Gender analysis in individual and household surveys

45. Gender analysis primarily identifies and addresses gender inequalities by acknowledging differences between and among women and men based on the unequal distribution of resources, opportunities, constraints, and power. However, even when gender data are produced in a survey, thorough analysis is not always conducted. Often, the utilization of gender data is limited by insufficient tabulation and dissemination of available information. Analysis of gender data can rely on different techniques, such as sex ratio, Gender Parity Index (GPI), gap analysis, correlations between gender and outcome variables, and trend analysis. Consequently, the effective use of gender data is hindered by challenges in translating and communicating statistics, leaving their potential underutilized in shaping gender-responsive policy discussions.[45]

46. Analyzing data by household composition (e.g., lone-person household, couple-only household, couple household with a child under six years, an extended family household with a child under six years) is also important in gender analysis because it provides critical insights into how resources, responsibilities, and opportunities are distributed within a household, which directly affects individual well-being and social outcomes. (See sample UN Women analyses by household composition for reference[46] [47] [48].)

47. Gender analysis also calls for the need to investigate intersectionality. Intersectionality highlights how gender intersects with other identity categories, such as ethnicity, ability, location, education, household income, and age, and can affect how people experience life. Some groups of people can experience multiple forms of marginalization. **Figure 12.1** demonstrates that each category could be analysed individually but should be nested with each other to shed light on the most marginalized groups and have targeted specific policies. (See sample on producing and using disaggregated gender statistics

[44] International Institute for Population Sciences (IIPS) and ICF. 2021. National Family Health Survey - 5 (NFHS-5), 2019-21. Accessed at https://dhsprogram.com/pubs/pdf/FR375/FR375_II.pdf

[45] World Bank. https://www.worldbank.org/en/topic/gender/brief/strengthening-gender-statistics

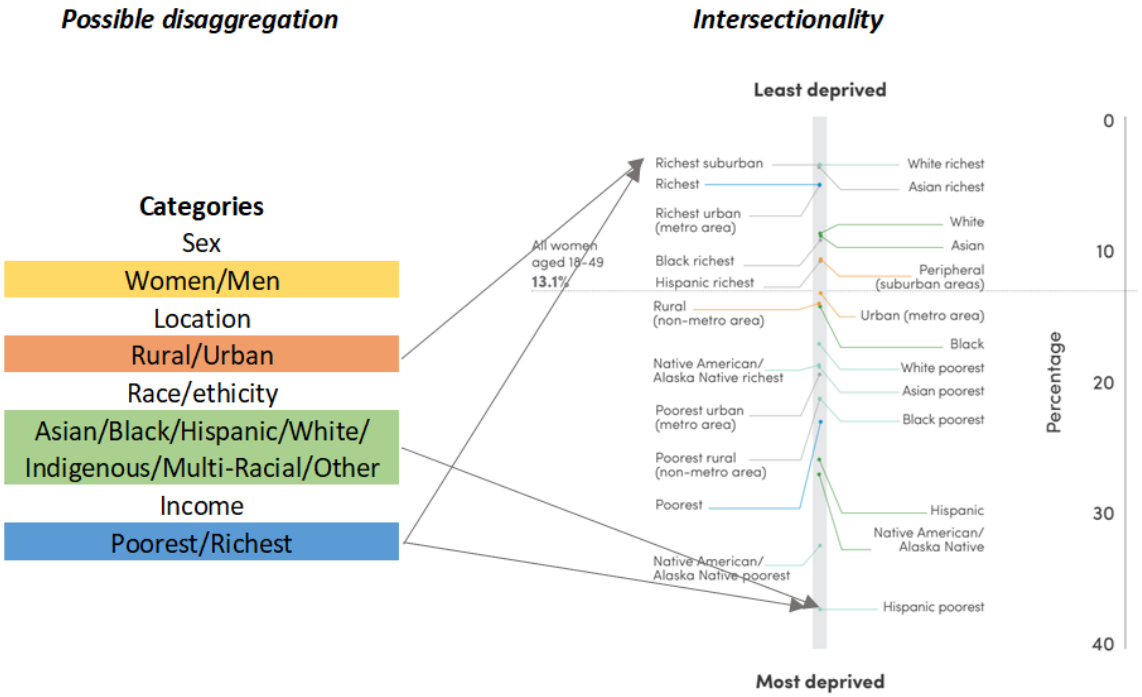
[46] UN Women. 2019. Progress of the World's Women 2019-2020: Families in a Changing World. https://www.unwomen.org/en/digital-library/publications/2019/06/progress-of-the-worlds-women-2019-2020

[47] ILO and UN Women. 2020. Spotlight on SDG8: The Impact of Marriage and Children on Labour Market Participation. https://data.unwomen.org/publications/spotlight-sdg8-impact-marriage-and-children-labour-market-participation

[48] Boudet, A.M.M., Bhatt, A., Azcona, G., Yoo, J., and Beegle, K. 2021. A Global View of Poverty, Gender, and Household Composition. World Bank Group Policy Research Working Paper.

from the Counted and Visible Toolkit[49] of UN Women and the Inter-Secretariat Working Group on Household Surveys for reference.)

Figure 12.1. Intersectionality analyses and an illustration using data from the United States on the proportion of women aged 18–49 who lack health insurance. Source: UN Women. 2018. Turning Promises into Action.[50]



48. Gender statistics cuts across many other fields of statistics to reflect the realities of women's and men's lives and policy issues relating to gender equality. The following issues include select statistical areas relevant to gender equality that are often prone to misinterpretation. This list of areas is not necessarily comprehensive. Still, it can provide a rough idea of some of the key issues that gender data users often encounter and how to deal with them to interpret gender data correctly.[51]

- Violence and crime data must always be interpreted carefully. As these estimates refer to very sensitive issues, they are consistently underreported.

[49] UN Women and the Inter-Secretariat Working Group on Household Surveys. 2021. https://data.unwomen.org/resources/counted-and-visible-toolkit

[50] UN Women. 2018. Turning Promises into Action. https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Publications/2018/SDG-report-Summary-Gender-equality-in-the-2030-Agenda-for-Sustainable-Development-2018-en.pdf

[51] UN Women and the UN Statistical Institute for Asia and the Pacific. Gender Statistics Training Curriculum. https://data.unwomen.org/resources/gender-statistics-training-curriculum

- Time use data is essential for advancing gender equality, but its analysis requires careful consideration. Specifically, accounting for simultaneous activities is crucial. When data only focuses on primary activities, tasks like supervisory care[52] are often overlooked, leading to undercounting women's time spent on these important activities. Additionally, analyzing time use data alongside sex disaggregation and integrating other variables—such as age (e.g., unpaid work as a key factor in the overrepresentation of young women in the "youth not in education or employment" category), location (e.g., increased unpaid care work due to activities like fetching water), income (e.g., the vicious cycle of income and time poverty), or access to care services—reveals the intersectional dynamics of time use

- Poverty rates are typically calculated at the household level. Poverty measures for women and men have traditionally been calculated by utilizing household measures of income or consumption and matching this information with the number of women and men living inside each household. However, such measures fail to capture intra-household inequalities. Monetary or otherwise, resources are often unequally distributed among household members. To capture accurate measures of individual poverty, separate assessments of income and/or expenditure at the individual level are necessary. In practice, few surveys ask for this level of information.

- The gender pay gap measures what women are paid relative to men. It is important to note that the data for this indicator should always be interpreted by occupation and level, considering the total time worked. Therefore, the indicator can be used to assess if equal pay is in place for equal work.

12.10. Disseminating survey results

49. Applying a gender lens to statistical processes aims to provide evidence of inequalities across social, economic, environmental, and political domains. Disseminating gender statistics requires tailored strategies to meet the needs of diverse users, including:

- Users unaware of the gender statistics they need.
- Users unsure where to find gender statistics in reports.
- Users who know how to find and use gender statistics but lack knowledge of tools.
- Users familiar with data tools but uncertain about interpreting gender statistics.

50. To address these challenges, a gender perspective should be integrated when building dissemination teams by including gender specialists as advisors; identifying target users by reaching out to women's organizations, media, policymakers, and support centers addressing gender issues; shaping key messages that may focus on survey-centric indicators (e.g., labour or household surveys) or monitoring-centric ones tied to specific goals (e.g., NPGEIs, SDGs); presenting data by using gender-sensitive visualization and language, as detailed in the *UN Women Counted and Visible Toolkit*[53]; timing releases, aligning with women-focused events like International Women's Day or Girl Child Day.

[52] For more information on how to measure and analyse supervisory care please see the UN Guide to Producing Statistics on Time Use (Box II.4 and IX).

[53] UN Women and the Inter-Secretariat Working Group on Household Surveys. 2021. https://data.unwomen.org/resources/counted-and-visible-toolkit

51. Monitoring dissemination effectiveness is crucial. This includes tracking whether materials reach target audiences and are used to drive positive change. User engagement can be measured through specific requests via websites, email, or phone.

52. Lastly, effective gender-sensitive dissemination requires intentional investment in human resources, ICT, and financial support, as emphasized in gender-responsive statistical budgeting indicated in Section 12.3.

CHAPTER 13
REFUGEES AND INTERNALLY DISPLACED PEOPLE

This is a placeholder page.

Chapter 13 will be completed in Spring 2025.

This page has been intentionally left blank.

CHAPTER 14
HOUSEHOLD SURVEYS AND EDUCATION

Manos Antoninis (UNESCO)
Silvia Montoya (UNESCO)

CHAPTER 14
HOUSEHOLD SURVEYS AND EDUCATION

14.1. Introduction

1. Household surveys are a data source for education statistics that complements or substitutes data from administrative records. For some indicators (e.g., adult literacy, adult participation in training), survey data are the only source. Even if they are not explicitly designed for education purposes, household, and other survey data have a number of advantages, which can be useful to understand education:

- *Policy application*: Surveys collect information on individual and household background characteristics that permit disaggregation of education indicators by sex, location, income or wealth, ethnicity, language and disability (as well as the intersection of these characteristics). Such information helps identify possible causes of observed social and economic outcomes, which can be used in policy design and provides insights into policy implementation. In contrast, disaggregation is challenging with administrative data.

- *Consistency*: Often two different data sources need to be combined to calculate key education indicators. At times, administrative data are inconsistent. For example, enrolment ratios have relied on the combination of enrolment counts from ministries of education and population data from national statistical offices. The two sources may not agree with each other, e.g. the population of students aged 10 may exceed the population of 10-year-olds. Surveys address this problem by providing internally consistent information on both components of population-based indicators.

- *Comprehensiveness*: Surveys collect information on some education indicators for which administrative data are not well suited. For example, surveys can assess the following: early childhood care and education or non-formal education and training (because it is simpler to ask service users than dispersed service providers); selected skills among adult population, such as ICT or literacy (which can only be assessed on a sample basis); and household education spending (as the household is the source of information).

- *Lifelong perspective*: Most discussions on household surveys and education tend to focus on attendance and attainment indicators. But the Sustainable Development Goal 4 framework gives a lifelong learning framework. Household surveys, if appropriately designed, can provide information on the following:
    - indicators:
        - education participation and attainment, e.g. out-of-school rate, completion rate
        - education environments, e.g. language used in school, bullying
        - education outcomes, e.g. literacy, self-reported digital skills
    - levels of education
    - modalities of education: formal, non-formal, and informal

2. The usefulness of household and other surveys is greater when they are part of a regular programme, which helps ensure the production of comparable statistics over time. Survey types relevant for education include income and expenditure surveys, labour force surveys, and multipurpose surveys, including international programmes such as the Demographic and Health Survey (DHS) and the Multiple Indicators

Cluster Survey (MICS). Past overviews of selected issues related to the use of household surveys in education statistics include Glewwe (2000), UIS et al. (2004), EPDC (2009), and UIS (2020).

3. Although researchers and international organizations have used household surveys extensively to analyse education issues, officials in ministries of education in many low- and middle-income countries have been less inclined to use them. Having historically relied on administrative records, many find it difficult to interpret survey findings especially when they appear to be inconsistent with administrative data (UIS and GEM Report, 2022). Equity in education, including the status of out-of-school populations, has also received insufficient policy emphasis. Partly as a result of that, there is relatively little involvement of education ministries in the design of national household surveys administered by national statistical agencies. Hence education questions sometimes do not capture the national education structure well, undermining analysis and cross-country comparisons.

4. This chapter highlights key considerations for basic and advanced questionnaire issues and analytical tasks when planning surveys to measure education indicators. It also places future methodological developments within the education data institutional structure and concludes with key takeaways.

14.2. Basic questionnaire design issues

5. In order to measure education indicators accurately and maximize comparability between surveys, it is important to add specific questions in a fixed sequence to the household roster section. The roster is the natural place for these basic questions to ensure that the education status of all household members is recorded. A household roster should include a minimum of two sets of questions:

[1] *Attainment* for all household members
A1. Has (name) ever attended education?
If yes:
A2. What was the highest education level (name) attended?
A3. Within that education level, what was the highest grade/year (name) completed?

[2] *Attendance* for some household members
B1. Is (name) attending education during the (20XX) school year?
If yes:
B2. What education level is (name) attending?
B3. What grade/year is (name) attending?

6. It is recommended that all surveys follow this structure and language. Yet even these seemingly straightforward questions need to address seven key challenges, described next.

14.2.1. Attainment/attendance: Level of education

7. ***Issues:*** Questionnaires tend to list the following education level options: pre-primary, primary, secondary and post-secondary education. Care needs to be exercised, as the definition of these levels may have changed over time, which is why it is important to also record the grade/years completed. Another problem is that some questionnaires include other descriptors (e.g. religious, non-formal) which do not correspond to a recognized level but to a different education modality.

8. ***Solution:*** Consult with ministries of education to ensure commonly understood national terms for education levels are used to minimize misreporting. It is recommended to create a new variable at the analysis stage that aligns nationally defined levels of education to the International Standard Classification of Education (ISCED) to facilitate international comparisons. Avoid conflating (formal) levels with (non-

formal) modalities as, in some countries, those who attend non-formal education may do so in addition to attending formal education (e.g. religious). If non-formal education is a common pathway in a country, develop additional questions to capture this phenomenon.

14.2.2. Attendance: Reference year

9. ***Issue:*** Some questionnaires ask whether an individual is 'currently' attending education. There is a risk of inaccuracy if a survey takes place during an academic holiday period and a risk of misclassification if a survey takes place over two academic years.

10. ***Solution:*** Specify the academic year to which the attendance question refers to prevent misleading responses. Record this information in the survey's background documentation.

14.2.3. Attendance: Reference age group

11. ***Issue:*** Education attendance questions tend to be asked of individuals aged five to 24 years. However, censoring of the target age group both to the left and to the right is limiting information. First, many children under five attend some form of education, often as part of an early childhood development or pre-primary education programme. Second, many education systems are trying to open more pathways into post-secondary education, and some bachelor programs are longer than five years.

12. ***Solution:*** Expand the age range of roster education questions to the group three to 29 years.

14.2.4. Attainment: Additional questions on graduation

13. ***Issue:*** Household survey questionnaires typically do not capture information on whether individuals completed an education level with the relevant certificate or degree. This is particularly important with respect to secondary education certificates, professional qualifications, and university degrees.

14. ***Solution:*** It is recommended that survey questionnaires ask respondents to specify whether they have acquired a certificate, qualification or degree at the end of the education cycle, following consultation with ministries of education.

A4. Did (name) receive a graduation certificate/qualification/degree from that education level?

14.2.5. Attendance: Additional questions on school type

15. ***Issue:*** Growing levels of participation in private education institutions is an important public policy issue. Data on the preschool, school, or university type attended can be used to analyse and explain trends. The challenge is to use a terminology that clearly distinguishes between institutions that do and do not charge fees, and between institutions that are and are not subsidized by the government.

16. ***Solution:*** It is recommended that survey questionnaires ask respondents to specify the type of education institution attended, following consultation with ministries of education, using a classification that is both easily understood by respondents and policy relevant.

B4. What type of education institution is (name) attending?

14.2.6. Attendance: Additional questions on early childhood care and education

17. ***Issue:*** Early childhood care and education is a diverse sector whose complexity is not easy to capture consistently across dimensions of duration, location, and funding due to a multiplicity of factors. here is a usually a distinction between services delivered to children up to and above age three.

18. ***Solution:*** It is recommended that survey questionnaires try to capture the diversity of services at this level with questions that relate to duration and location for children aged three and above, for whom it is reported that they participated in early childhood development or pre-primary education.

> B5. Does (name) receive early childhood development services in a centre?
> If yes:
> B6. How many hours per week does (name) normally spend at the centre?

14.2.7. Attendance: Additional questions on post-secondary education

19. ***Issue:*** Household survey questionnaires do not tend to provide sufficient information on the post-secondary education course attended.

20. ***Solution:*** In addition to the question recommended above on graduation, it is recommended that survey questionnaires try to capture the duration and modality of post-secondary education, for those for whom it is reported that they participated in post-secondary education.

> B7. How many years does it take for the post-secondary course (name) is attending to be completed?
> B8. Does (name) attend this course from a distance?
> B9. Does (name) attend this course full time?

14.3. Advanced questionnaire design issues

21. In addition to attendance and attainment, harmonization of questionnaires is advisable in a range of other issues. This chapter focuses on three topics that deserve attention: technical and vocational education and training; skills; and education expenditure.

14.3.1. Attendance: Additional questions on technical and vocational education and training

22. ***Issue:*** Technical and vocational education and training is the most diverse sector whose complexity in terms of duration, location, funding, and purpose is very difficult to capture, for example in global SDG indicator **4.3.1**.

23. ***Solution:*** It is recommended that (labour force) survey questionnaires try to capture the diversity of training with questions that relate to purpose, duration, and modality for adults aged 25 to 64.

> C11. In the last 12 months, did (name) participate in education and training for professional reasons?
> If yes:
> C12. How many hours did (name) spend in education and training in the past 12 months?
> C13. Did (name) attend receive this education and training in an education institution?
> C14. Did (name) attend receive this education and training at the workplace?
>
> C21. In the last 12 months, did {name) participate in education and training for personal reasons?
> If yes:
> C22. How many hours did (name) spend in education and training?

14.3.2. Skills

24. Collecting information on education outcomes through surveys is complex and costly. A full treatment of the relevant issues is outside the scope of this overview chapter, but three issues are addressed briefly below.

14.3.2.1. Adult literacy rate

25. ***Issue:*** Information on adult literacy has historically relied on self-reported answers to questions such as 'can you read', which have not proven to be reliable or comparable (e.g. Nath, 2007). Direct assessment is needed to ensure high-quality measurement of SDG global indicator **4.6.1**.

26. ***Solution:*** All household surveys, at least in low- and middle-income countries, should administer a direct reading test to the adult population, even to those who have attended secondary education, such as the one used by DHS which relies on cards with short four- to six-word sentences.

14.3.2.2. Foundational literacy and numeracy

27. ***Issue:*** Although a school-based learning assessment and the leadership of the education ministry should be the first choice for measuring SDG global indicator **4.1.1a** (literacy and numeracy skills by the end of grade 2 or 3), a household-based assessment is possible for low- and lower-middle-income countries as long as it meets reporting criteria.

28. ***Solution:*** Household survey designers can use the Assessment for Minimum Learning Proficiency household module, which assesses foundational literacy and numeracy skills.

14.3.2.3. ICT skills

29. ***Issue:*** Global SDG indicator **4.4.1** on information and communication technology (ICT) skills consists of self-reporting on nine ICT-related activities in the three months preceding the survey. It is an indicator in need of constant evolution given the rapid technological changes.

30. ***Solution:*** Household survey designers should consult with the Expert Group on ICT Household Indicators, which coordinates efforts on the development of this indicator (ITU, 2024) within the conceptual framework of the Digital Competence Framework for Citizens (Digcomp) (European Commission, 2024).

14.3.3. Education expenditure

31. ***Issue:*** Education expenditure questions are a standard feature of household income and expenditure (or budget) surveys. However, there remain important differences between national surveys in the extent of detail and in the way these questions are administered: expenditure categories (e.g., fees), separating expenditure paid to institutions from other expenditure, questions per individual or for the entire household, recall periods, distinction by type of education institution attended etc.

32. ***Solution:*** A standard set of questions on education expenditure has been proposed as reference for survey practitioners (Oseni et al., 2018). The most important recommendation is that surveys should collect expenditure information for each household member and each item to improve accuracy.

14.4. Data integration

33. Education statistics need to use multiple data sources or types of data sources in the estimation of headline indicators. Other sectors have previously faced such challenges. For example, the need to use multiple surveys with different methodologies (as well as to address data gaps) to estimate wasting and stunting rates led to the establishment of the Joint Child Malnutrition Estimates inter-agency group (UNICEF et al., 2023). The need to use administrative and survey data sources to estimate health indicators

led to the establishment of the UN Inter-agency Group for Child Mortality Estimation (Alkema and New, 2014) and the UN Maternal Mortality Estimation Interagency Group (Alkema et al., 2016), which developed estimation models.

34. Two Bayesian hierarchical models, inspired by the approach used to estimate health indicators, have been adapted to the education context:

- The model for the completion rate, which is SDG global indicator 4.1.2, uses multiple household surveys (Dharamshi et al., 2022), which if examined in isolation could – due to differences in methodologies, sampling, objectives or circumstances at the point of data collection – produce results that are not fully comparable.
- The model for the out-of-school rate (UIS and GEM Report, 2022) combines household survey and administrative data, making efficient use of both sources.

35. As more indicators are estimated drawing on household survey data, new challenges emerge that need to be addressed (UIS, 2024a):

- Ensure best practice in reporting estimates based on multiple data sources
- Ensure country participation and ownership in the generation of such estimates
- Ensure consistency between various estimates (e.g. out-of-school and completion rates)

14.5. Institutional architecture for education data and statistics

36. The Conference on Education Data and Statistics is the apex body with the mandate to communicate, discuss, and reach consensus on key issues regarding concepts, definitions, methodologies, and operational aspects of education indicator measurement in the form of recommendations and guidelines for adoption as international standards to improve comparability. It takes place every three years. Each country is invited to send three members to the Conference, one of which is a national statistical office representative. The implementation of the Conference decisions is the responsibility of the Education Data and Statistics Commission, which consists of 28 member states at the time of writing. The Commission has five working groups, one of which focuses on household surveys. The scope of its work is to address all issues raised in this annex (UIS, 2024b).

14.6. Key takeaways

37. Countries need to:

- Develop capacity to use education questions in household surveys for policy planning.
- Design education questions in household surveys to take into account the following:
  - Adopt a lifelong perspective to education, therefore improving the way questions are asked on early childhood, post-secondary and adult education
  - Use the more flexible approaches through which education is accessed, therefore capturing distant and non-formal modalities
- Ensure questionnaires are simple, keeping filter questions to a minimum, and aligned with good international practice

This page has been intentionally left blank.

CHAPTER 15
MEASURING GOVERNANCE, PUBLIC SAFETY, JUSTICE, AND CORRUPTION

Alexandre Wilde (UNDP Global Policy Centre for Governance)
Arvinn Gadgil (UNDP Global Policy Centre for Governance)
Fatma Usheva (UNDP Global Policy Centre for Governance)
Mariana Neves (UNDP Global Policy Centre for Governance)
UNODC Research and Trend Analysis Branch

CHAPTER 15
MEASURING GOVERNANCE, PUBLIC SAFETY, JUSTICE, AND CORRUPTION

15.1. Introduction

1. Governance is the cornerstone of development for all countries because it establishes the framework for decision-making, public administration, and the distribution of power and resources. The 2030 Agenda recognises this critical role by including indicators on peaceful, just, and strong institutions under Sustainable Development Goal 16, affirming that good governance is essential to achieving development in health, education, climate, and beyond.[1] Without good governance, aspirations for sustainable development remain unattainable.

2. To harmonise and standardise the concept of governance for statistical purposes, the United Nations Statistical Commission (UNSC) has included governance into the Classification of Statistical Activities[2], addressing the plethora of interpretations and definitions surrounding the concept. The governance statistics domain is divided into eight areas of statistical activity[3]: (1) non-discrimination and equality, (2) participation, (3) openness, (4) access to and quality of justice, (5) responsiveness, (6) absence of corruption, (7) trust, and (8) safety and security. These eight areas are currently at varying stages of methodological development. For example safety and security[4] (along with its sub-activity crime and criminal justice[5]), absence of corruption, responsiveness (e.g. service delivery efficacy and external political efficacy), and access to and quality of justice are in a more advanced state with some approved methodological standards and guidelines already in place.[6] In contrast, areas such as non-discrimination and equality, openness, trust in institutions as well as participation in political and public affairs are in the methodological development phase.

3. As with other domains, governance statistics rely on both administrative data and representative household surveys, and there is a growing interest and use of other sources such as citizen data.[7]

[1] For further information on the 2030 agenda, refer to General Assembly resolution 70/1.

[2] For further information on the classification, refer to the Classification of Statistical Activities 2.0 (E/2023/24-E/CN.3/2023/37).

[3] For further information on the concept of governance and respective subdomains, please refer to the Handbook on Governance Statistics (E/2020/24-E/CN.3/2020/37).

[4] Covers statistical activities related to crime, victimization, violence, perceptions of safety, and the quality of criminal justice institutions, among other activities.

[5] Covers statistical activities related to crime, criminal justice systems, and victimization.

[6] A list of methodological standards, statistical classifications, and technical guidelines related to the topics covered in this chapter is provided at the end of the chapter.

[7] "Citizen data" as currently defined in the Copenhagen Framework on Citizen Data are data originating from initiatives where citizens either initiate or are sufficiently engaged, at the minimum, in the design and/or collection stages of the data value chain, irrespective of whether these data are integrated into official statistics. The Framework was endorsed by the United Nations Statistical Commission. While citizen data is beyond the scope of this chapter, there is a growing interest in using citizen data in various statistical domains including governance. One

Administrative data are essential for producing governance indicators and rely on data collected by government agencies and institutions in their day-to-day operations. Such data mainly reflect operational realities of governance systems, including institutional capacity, processes, resources, and infrastructures. However, administrative data lack the human centred perspective, as the information is generally more of administrative nature. Household surveys, on the other hand, play an equally important role by filling in the gaps left by administrative data and are the primary source for collecting data on indicators that measure people's experiences, behaviours, perceptions, and attitudes. For example, crime victimization surveys or survey modules measuring access to civil justice reveal critical information absent from administrative records. These tools capture incidents that go unrecorded by the police or other dispute resolution mechanisms, and they reveal the "dark figure" of crime[8] and the "justice gap",[9] offering an understanding of the prevalence of victimization and legal needs among the population, as well as public attitudes toward crime, perceptions of safety and justice. This people-centred approach emphasises the need for data from the perspective of the population and victims, rather than solely relying on the perspective of the government service provider.

4. Household surveys are also essential for assessing and monitoring the relationship between the state and the population, ensuring governance is inclusive, transparent, and accountable. In governance statistics, there is a frequent use of both objective and subjective data.[10] Objective data comprises experiences of respondents based on real events, actual experiences, and behaviours. Subjective data, represents the respondent's perceptions, opinions, attitudes, and feelings regarding a particular issue. While objective data is more established, subjective data, collected through household surveys, is equally important for a comprehensive understanding of a phenomenon.

5. Governance topics such as justice, victimization, and corruption present unique challenges ranging from underreported, hidden, and stigmatised experiences to the complexity of understanding public opinions of government performance. Practitioners conducting household surveys in these fields should consider several key factors. This chapter offers guidance, outlines key considerations, and provides useful references for implementing surveys or survey modules for governance statistics. This chapter focuses on several domains with more established methodological standards, including access to and quality of justice, responsiveness, absence of corruption, trust, and safety and security.

15.2. Needs, scope, and business case

6. By definition, governance is a multi-sectoral domain, connecting areas such as education, health, justice, internal affairs, among several others. Producing governance statistics, therefore, requires a coordinated approach across all relevant data producers and data users. A broad consultative process with diverse stakeholders is fundamental when identifying needs and defining the scope of data collections. This

example is a citizen data pilot project implemented by Ghana Statistical Service to understand and measure satisfaction with public service delivery in Ghana.

[8] A sizable portion of criminal events are never reported to the police and are therefore not included in police or any other statistics. This unknown number is often referred to as the 'dark figure' of crime. For more detailed discussion of the concept, including its measurement, see UNODC-UNECE, Manual on Victimization Surveys (2010).

[9] The "justice gap" can be understood as the number of people who have at least one unmet justice need. See "Measuring the Justice Gap: A People-Centered Assessment of Unmet Justice Needs Around the World", The World Justice Project, 2019.

[10] Praia Group on Governance Statistics, *Handbook on Governance Statistics* (2020)

comprehensive approach ensures inclusivity, relevance, and efficiency, while avoiding duplication in data collection activities.

7. During the initial needs assessment, stakeholders should review established international standards and tools, including the Sustainable Development Goal 16+ indicators, manuals on corruption and victimization surveys, and relevant statistical classifications such as the International Classification of Crime for Statistical Purposes (ICCS). (For a comprehensive list and references to international standards, tools, and statistical classifications relevant to governance data, please refer to the end of this chapter.) Adopting these methodologies ensures the survey is designed using impartial, well-tested tools that meet both national and international data needs.

8. However, while these international standards provide critical guidance, the scope of governance surveys should reflect national priorities and needs. Therefore, broad consultation with all relevant stakeholders is also essential to ensure the survey (a) collects data that is relevant to the national context, (b) can be used by different stakeholders, and (c) has national ownership. During this stage, stakeholders should agree on key impartial indicators to be measured and on the methodology to be used that can ensure this impartiality. Agreeing on these aspects before data collection can help avoid bias or lack of trust in the results. This is especially crucial when addressing politically sensitive topics such as corruption, public trust in institutions, government responsiveness, or other indicators that measure the functioning and effectiveness of public institutions.

9. Conducting dedicated surveys to measure governance, safety, and related issues in greater detail may not always be feasible due to resource constraints, logistical challenges, or political sensitivities. In such cases, such challenges can be addressed by integrating some indicators that are considered national priorities into existing survey instruments that already cover related topics, such as household surveys on living conditions or social attitudes. While this approach may not cover all governance topics or capture the same level of detail as a standalone survey, it allows for efficient data collection in resource constrained environments – leveraging established infrastructure and ensuring that key metrics are captured.

15.3. Survey management

10. Certain governance topics, such as corruption or trust in institutions (including government), are highly sensitive not only for the respondents but for the institutions that are being assessed. Ensuring the quality, credibility, and reliability of data while preserving national ownership in politically sensitive surveys requires careful planning and risk mitigation. It is essential to involve a wide spectrum of stakeholders in the survey process to enhance the legitimacy and quality of data and reinforce the data's objectivity. This goal can be accomplished through the establishment of formal technical and/or steering committees comprising representatives from national development agencies, line ministries, national government departments and institutions, civil society, and international organizations. These committees can oversee the entire survey implementation process from inception to dissemination, and provide advice to the technical team and management team on the topics of their domain, thus promoting transparency, independence, and trust in the process.

15.4. Questionnaire

11. Surveys on governance justice, corruption, and victimization, aim to measure experiences, opinions, and attitudes that may be politically sensitive, socially undesirable, or even illegal. In such cases, there is heightened risk that respondents may be reluctant to provide honest answers or may give politically safe answers to avoid perceived risks. This issue is particularly pronounced if respondents believe they are responding directly to a government agency (and certain answers could carry negative repercussions).

Research has shown that social desirability bias is more prevalent for sensitive questions than non-sensitive questions.[11] As such, soliciting honest and accurate responses requires a range of approaches. Key solutions include building strong rapport with respondents, assuring them of their anonymity and confidentiality, and wording questions in a neutral, non-leading manner to minimise bias. These practices are foundational to the design and implementation of sensitive surveys.

12. For cases where respondents may not provide truthful answers to sensitive questions, different techniques such as list experiments can be effective. A list experiment is a survey technique that allows respondents to indicate their experiences or opinions without directly answering sensitive questions.[12] Instead, respondents are presented with a list of statements and are asked to indicate how many of them are true, without specifying which ones. This indirect method allows participants to share sensitive information, can help reduce social desirability bias and encourage more honest responses. While not a new technique, list experiments have proven effective in improving the quality of data on sensitive issues and could be considered for surveys addressing topics such corruption, victimization or other politically sensitive topics within the governance statistics domain.

13. Another challenge within questionnaire design is the use of complex or technical concepts and legal terminologies that may confuse respondents and lead to reduced accuracy and comparability of answers. For example, respondents may not fully understand what constitutes certain forms of violence or discrimination or may be unfamiliar with legal terminology associated with civil and criminal justice. Respondent understanding of all these concepts may vary widely based on cultural, educational, or social factors, and such variations can affect the accuracy and comparability of data across different countries and even within the same country over time.

14. One solution to this challenge is to provide respondents with concrete examples that clarify these concepts without using overly technical language. For instance, instead of asking respondents if they have experienced “civil disputes”, multiple examples in the form of questions may specify what civils disputes are such as “Have you personally experienced problems with land, or buying or selling property?” or “Have you personally experienced family issues such as divorce, child support, child custody or will?”. Since it is impractical to list all potential examples, the existing list of examples for the measurement of these concepts could be drafted based on a pilot with inputs from relevant experts and institutions, or based on a more detailed data collection which would allow to extract the most frequent disputes and group them into broader types of disputes for future surveys. Moreover, it is also important to strike a balance between providing sufficient examples, questionnaire length, and respondent burden. This balance can be tested and evaluated through cognitive testing and pilot surveys. Cognitive testing ensures that questions are understood as intended and are culturally appropriate, while pilot surveys test the instrument in a real-world setting on a small number of respondents, allowing for refinement of questions, flow, and methodology before full-scale implementation.

15. Beyond comparability of answers from the same survey instrument, aligning questionnaire design with global standards, statistical classifications, and methodologies – including the SDG indicators methodology – ensures coherence between survey data and other data sources. This alignment facilitates more reliable comparisons and analyses across datasets. For instance, designing questions on criminal

[11] Krumpal, Ivar, Determinants of social desirability bias in sensitive surveys: a literature review, Quality and Quantity, 2011

[12] Blair, Graeme, and Kosuke Imai. 2012. “Statistical Analysis of List Experiments.” Political Analysis 20(Winter): 47–77.

offenses based on definitions from the ICCS allows survey data collected from corruption or victimization surveys to be compared with administrative records from the criminal justice system. Such harmonizations strengthen the utility of both data for policy and decision-making.

16. Another important consideration in questionnaire design is the selection of a reference period. The choice of reference period depends on the survey's objective and the frequency of the events being measured. In most international surveys on corruption and victimization, a 12-month reference period is commonly used. This reference period strikes a balance between capturing sufficient data for analysis and minimizing possible bias in recalling events. Reference periods should always be consistent to mitigate problems related to non-recall and mis-recall. Non-recall usually depends on the memory decay of respondents, while mis-recall is mainly based on the telescoping effect where respondents have difficulty accurately locating events within the appropriate reference period.[13] Techniques to handle mis-recall include time anchoring, where significant public events are used as anchors (such as national holidays or major news events). Some offences such as intimate partner or workplace violence present unique challenges as they may lack clear beginnings or ends. As a result, respondents may report repeated incidents multiple times during the reference period. This challenge can be addressed by capping the number of incidents recorded, which, while potentially undercounting the extent of crime victimization, ensures more consistent comparisons and reduces the impact of outliers.

17. To support NSOs and other practitioners, the *Manual on Victimization Surveys* and the *Manual on Corruption Surveys* include details of all aspects of preparing a questionnaire and data collection organization for crime-victimization and corruption surveys. Examples of rigorously tested tools for collecting governance and victimization indicators are the *SDG16 Survey Initiative* and the Latin America and the Caribbean Crime Victimization Survey Initiative (LACSI). Readers are encouraged to consult these resources referenced at the end of this chapter for comprehensive insights into survey design and implementation of governance, justice, victimization and corruption surveys.

15.5. Frame and sampling

18. Hard-to-reach or hidden populations are key target groups for surveys on victimization and discrimination. Targeted sampling or oversampling is recommended for those groups, but the lack of available sampling frames can necessitate the development and use of innovative methods to reach them. One such method is the Network Scale-Up Method (NSUM)[14] which allows for the estimation of prevalence indicators among hidden populations by analysing sampled social network data. The NSUM involves including a set of questions in a general population survey to estimate the size of the respondent's personal network. These data are then used to estimate the prevalence of the hidden phenomenon being measured. For example, the NSUM has been piloted to estimate the prevalence of trafficking in persons, as victims of trafficking tend to be more hidden than other victims of crime and more difficult to access due to low self-identification.[15]

[13] Please see Manual on Victimization Surveys and *SDG 16 Survey Initiative Manual* for further information on recall, telescoping and time anchoring.

[14] More information on Network Scale-Up Method available in: Proposed Utilization of the Network Scale-Up Method to Estimate the Prevalence of Trafficked Persons, Janie F. Shelton, https://www.unodc.org/documents/data-and-analysis/Forum/Forum_2015/15-00463_forum_ebook_E.pdf

[15] For more information, see *Fiji National Trafficking in Persons Prevalence Survey*

19. The NSUM-based approach can also be used in household surveys to estimate the number of victims of intentional homicides.[16] In addition, a survey module has also been developed for inclusion in general population surveys allowing for the estimation of homicide rates by sex, region, situational context, and the mechanism of killing.[17]

15.6. Data collection

20. Collecting data for governance and other sensitive topics such as corruption and victimization requires specialised field protocols and well-designed interviewer trainings equipping them with the necessary tools and skills to address all the challenges associated with sensitive questions.

21. One critical factor in data collection is the choice of survey mode. While face-to-face interviewing has been the primary mode of data collection, the growing use of information and communication technologies have made alternative modes increasingly more attractive and feasible. However, the efficacy and efficiency of different survey modes will vary significantly depending on a number of factors which should be carefully assessed prior to making a decision. For example, self-administered surveys produce lower social desirability bias compared to interviewer face-to-face surveys.[18] However, their feasibility of implementation depends on country-specific factors such as cultural norms, literacy rates, target population, infrastructure, etc. Similarly, computer-assisted telephone interviewing (CATI) has gained prominence especially since the onset of the COVID-19 pandemic which significantly disrupted regular data collection processes.[19] While CATI offers numerous advantages, its use for sensitive topics - such as corruption, victimization, opinions on system responsiveness, etc. – requires careful evaluation. Respondents may be reluctant to disclose truthful answers over the phone due to privacy concerns or potential misuse of information.

22. When implementing governance, justice, victimization, and corruption surveys, practitioners should evaluate some key considerations to determine the most appropriate survey mode:

- Assess response rates for similar surveys conducted in the country: response rates might vary greatly over time or between surveys, but such an assessment could provide an indication and allow for an informed decision.
- When collecting data on very rare events, such as some types of crimes, it is advised to avoid data collection modes that already have an inherently very low response rate in that specific country/context.
- The need to assess safety before and during interview: sensitive questions require additional safeguards during the interview which are not possible using CAWI and CATI. For instance, in

[16] McCormick, T.H. and Zheng, T. (2012) 'Latent demographic profile estimation in hard-to-reach groups', The Annals of Applied Statistics, 6(4). doi:10.1214/12-aoas569.

[17] See UNODC (2023), Global Study on Homicide 2023, page 98.

[18] Praia Group on Governance Statistics, *Handbook on Governance Statistics* (2020)

[19] Sydney Gourlay, Talip Kilic, Antonio Martuscelli, Philip Wollburg, Alberto Zezza, Viewpoint: High-frequency phone surveys on COVID-19: Good practices, open questions, Food Policy, Volume 105, 2021.

the case of questions on direct experience with intimate partner violence for which privacy is of paramount importance.

23. The existing protocols and procedures for data privacy and their exemptions: National data protection protocol, capacity, and infrastructure should be taken into consideration before choosing the data collection mode. For instance, the data collection of sexual orientation in a country that penalises certain orientations needs to be carefully considered.

24. In general, CAPI has demonstrated to perform well for surveys on the topics covered in this chapter, but the choice of mode should depend on the scope of each specific survey, budget, technical capacity, infrastructures, national legislations and protocols, implementation period, public acceptability, and response burden, among others.

25. Apart from survey modes, for politically sensitive surveys where results could be obscured for political or non-statistical purposes, a robust quality assurance framework is critical to ensure data integrity, accuracy, reliability, and dissemination. For topics such as corruption, involving an independent third party to oversee the quality of survey activities and outputs has proven successful. These independent third-party monitors ensure the quality of survey activities and survey outputs at all stages of the survey implementation, from the quality of survey instrument and its translation into local languages, to all other survey processes – including training of interviewers, logistical arrangements and organization of field operations, technical aspects of data capture, etc. – up to the assessment of data quality. Such third-party external quality assurance mechanisms have been adopted, and have been especially useful, during the implementation of corruption surveys. For example, an independent party was contracted for quality control during the third survey on corruption in Nigeria.[20] The independent third-party can also be a national institution with a mandate and capacity to monitor or to assess the implementation of the survey, such as an independent oversight mechanism or institution.

15.7. Dissemination

26. The dissemination of survey results covering statistics on governance, justice, victimization, or corruption requires careful management to ensure transparency, trust, and broad stakeholder ownership. Moreover, data on politically sensitive issues are subject to intense scrutiny from both the authorities and the general public. Given the political nature of governance statistics, additional steps are needed to ensure that the data are available to all stakeholders and accurately reflects the reality. This approach is essential to reinforce the legitimacy and credibility of the results.

27. A clear dissemination protocol should be established to ensure that finalised survey products are shared widely and responsibly. This includes setting guidelines on who will receive the results, when results will be published, and through which channels. Transparency in this process builds confidence in the data and ensures that all stakeholders, from government officials to civil society, have equal access to the information. NSOs play a crucial role in ensuring the impartiality and credibility of data and analysis. To maintain public trust and integrity, it is essential that NSOs adhere to the Fundamental Principles of Official Statistics (FPOS) (see Chapter 1), which specify 10 principles that guarantee professional independence, impartiality, confidentiality, transparency, and the quality of their statistical outputs.

28. To build and maintain trust in the national statistical system, it is crucial that the publication of politically sensitive data is guaranteed, regardless of the political context or the nature of the results.

[20] UNODC, *Corruption in Nigeria: Patterns and Trends* (United Nations publication, 2024).

Upholding the independence of the statistical office and adhering to established legal frameworks can ensure that the results of such surveys are disseminated impartially, reinforcing public trust in governance data and promoting evidence-based policy decisions.

15.8. International standards and classifications related to governance

Praia Group on Governance Statistics, *Handbook on Governance Statistics* (2020)

UNDP, UNODC and OHCHR, *SDG16 Survey Initiative* (2022)

UNODC, *International Classification of Crime for Statistical Purposes* (2015)

UNODC, UNDP and the UNODC-INEGI Center of Excellence in Statistical Information on Government, Crime, Victimization and Justice, *Manual on Corruption Surveys* (2018)

UNODC-UNECE, *Manual on Victimization Surveys* (2010)

UNODC, IDB, UNDP, OAS and the UNODC-INEGI Center of Excellence in Statistical Information on Government, Latin America and the Caribbean Crime Victimization Survey Initiative (LACSI) (2021)

This page has been intentionally left blank.

CHAPTER 16
LABOUR STATISTICS

Lara Badre (International Labour Organization)
Antonio R. Discenza (International Labour Organization)
Michael Frosch (International Labour Organization)
Rosina Gammarano (International Labour Organization)
Vladimir Ganta (International Labour Organization)
Kieran Walsh (International Labour Organization)
Samantha Watson (International Labour Organization)

CHAPTER 16
LABOUR STATISTICS

16.1. Background

1. While labour related statistics can be generated from multiple sources, a core source of data should be a dedicated labour force survey (LFS), designed to apply the latest international statistical standards. This chapter describes the key statistical standards and some important methodological considerations for the implementation of a labour force survey. The guidance would also be relevant to other household surveys seeking to measure labour related topics.

16.1.1. The International Labour Organization (ILO) mandate and the ICLS

2. The ILO is the international organization that primarily supports the development and implementation of statistical standards on labour. Statistical standards are adopted through the International Conference of Labour Statisticians (ICLS), which takes place every five years. These standards have covered a wide range of topics over time.

3. Support for implementation of statistical standards is provided through a mix of guidance development, capacity building, and technical assistance with the support of a network of regional labour statisticians. All published guidance is available through the ILOSTAT website, including a dedicated page on standards and methods devoted to the wide range of standards and related guidance to support those seeking to implement a LFS, or other household surveys including the measurement of labour. The ILOSTAT website also hosts the ILO's database of labour related statistics.

4. As of 2024, three resolutions are of particular note to inform the design of a labour force survey. They are:

- Resolution I, 19th ICLS – 2013 – resolution concerning statistics of work employment and labour underutilization: This resolution defines the overarching concept of work, different forms of work, including employment and multiple concepts of labour underutilization including unemployment
- Resolution I, 20th ICLS – 2018 – resolution concerning statistics on work relationships – establishing the latest version of the International Classification of Status in employment (ICSE-18) and a classification of status at work (ICSAW-18)
- Resolution I, 21st ICLS – 2023 – resolution concerning statistics on the informal economy – establishing a new framework for statistics on informality including definitions for key concepts such as informal employment and the informal sector

5. These three standards are interrelated by design and in implementation, as the later standards build on the definitions of the earlier standards.

6. Those designing a LFS, or other survey covering labour related topics, should review the definitions contained within the standards, as well as available model questionnaires (discussed further below). In addition to those described above, currently applicable standards cover topics such as the International Standard Classification of Occupations (ISCO), working time, child labour, labour migration, employment related income, forced labour, cooperatives, qualifications and skills mismatches and various others, all of which can be found at the ICLS documents page of the ILOSTAT website.

16.2. Key methodological issues for labour force surveys

16.2.1. Questionnaire design

7. The ILO has developed and published questionnaire content, designed to apply the latest statistical standards. The broad principles underlying the questionnaire content are:

- ***Efficient and flexible*** content with the possibility to cover different topics at different points in time or depending on national priorities. This principle is achieved through a modular approach whereby specified blocks of questions can easily be added or removed without requiring further redesign of the parent questionnaire
- ***Comprehensive coverage of key statistical standards*** reflecting the content needed to apply the agreed criteria and definitions
- ***Evidence-based development built*** on *extensive testing in different countries*
- ***Guidance and tools incorporate CAPI and PAPI*** versions with support available for CSPro, and for the main LFS content, also Survey Solutions

8. While the published content has been extensively tested, it remains recommended that countries also plan for national testing in advance of implementation, with technical assistance available from ILO to support this process. It is also recommended that translation to other languages is carefully implemented. It is possible to develop alternative questionnaire content to apply the standards, but the ILO strongly recommends that those doing so make reference to the published model questionnaires and related guidance, and comprehensively test the new questionnaires. The testing undertaken by the ILO has demonstrated (as with many other studies) that measurement is very sensitive to questionnaire content and casual/part-time employment are particularly subject to undercount, typically with a greater impact on estimates of women's employment than men's.

9. The 19th ICLS standards included a forms of work framework defining multiple forms of work. The main LFS questionnaire covers the measurement of employment in great depth. Guidance on the measurement of additional forms of work from the forms of work framework is described briefly below.

16.2.1.1. Measuring volunteer work

10. Volunteer work is an unpaid form of work in which important numbers of persons engage either through organizations or directly to maintain/improve wellbeing of other individuals and entire communities. The defining characteristics of volunteer work are that it is unpaid, non-compulsory, and the goods or services produced are for persons outside the producer's household or family. To make this work visible and recognise its developmental potential, the ILO recommends producing estimates of numbers of volunteers and hours they work using the latest international standards and tools described in the volunteer work measurement guide.

16.2.1.2. Measuring own-use production of goods

11. Own-use production of goods (OPG) includes activities to produce goods intended mainly for final use of the producer, their household, or family. In countries where agricultural activities are prevalent, it is important to cover own-use production of foodstuff regularly to provide comprehensive figures about the agriculture sector and enable more policy-relevant analysis. Own-use production of other goods (such as making clothing, fetching water, etc.) are also important to measure, but may be measured less frequently depending on national context and interest. ILO model questionnaires include content required to measure

different types of goods produced through OPG. A subgroup of OPG of particular note is subsistence foodstuff production and the 19th ICLS standards recommend the publication of an indicator called the subsistence foodstuff producer rate – in countries where this is prevalent.

16.2.1.3. Measuring own-use provision of services

12. Own-use provision of services (OPS) work is a relatively recent addition to the scope of official labour statistics, first being defined in Resolution I of the 19th ICLS. OPS work includes activities such as cooking, cleaning, household management and maintenance, pet-care, childcare and education, care of adults, and related travel and transportation[1]. The category of "unpaid domestic and care work", as specified for measurement in the Sustainable Development Goals, is identical with OPS work[2]. OPS work is characterised by the intended destination of the provided service, where that is "mainly for final use…or final consumption by household members, or by family members living in other households"[3]. In 2023, the ILO published add-on modules and guidance for national LFS to support the comprehensive measurement of OPS work and total (paid and unpaid) work through a single data source. In addition to adding questionnaire content, the coverage of OPS through the LFS has various additional methodological implications including for sampling and fieldwork practices – all of which is discussed in the published guide (linked above).

16.2.1.4. Measuring labour in household surveys other than LFS

13. The statistical standards can also be applied in other household surveys such as household income and expenditure surveys or general multi-topic surveys. In this case the same minimum requirements to apply the standards are still relevant, albeit questionnaire content may be more limited. In addition to the LFS the ILO has produced guidance specific to the population and housing census.

16.2.2. Sampling

14. Being the main source of official labour statistics on all forms of work from the labour supply side, the LFS produces a large variety of labour market indicators, including cross-sectional data, annual averages, quarterly net changes, annual trends, etc. Given the survey's high policy relevance, it is essential to monitor trends and seasonal variations effectively. Below are some key considerations for optimizing the sampling design.

15. To ensure relevance to users and stakeholders, data collection should occur regularly and on a sub-annual basis where resources allow, with consistent frequency year after year. In each survey round, estimates should reflect the average situation during the specific reference periods (e.g., a month, a quarter, a season, or a year). Therefore, the sample should be uniformly spread across weeks, months, and seasons. Within each of these periods, the sample must represent all the different geographical areas. Where continuous data collection is not possible (or not considered necessary), other approaches can be applied, for example picking one reference week per month. However, in this case the choice of the reference week is crucial for the quality of the results – with it being necessary to select a typical or representative week and avoiding weeks with particular peaks or troughs in activity (e.g., a holiday week).

---

[1] ILO (2013) Resolution I of the 19th ICLS, page 9, para 22(c)

[2] SDG indicator metadata [indicator 5.4.1 "the proportion of time spent in a day on unpaid domestic and care work by men and women"], 2023, United Nations Statistical Division, available here. Last accessed 18/07/2024

[3] Resolution I of the 19th ICLS, page 9, para 22(a)

16. LFSs typically use a two-stage stratified sample design (see Chapter 5). As specified in the relevant standards, all household members of working age (often 15 years or over) are generally interviewed, although basic demographic information should also be collected for all household members to allow analysis of labour indicators by household composition.

17. In case of LFSs carried out regularly, it is possible to benefit from repeat visits to the same households on a rotational basis – also referred to as sample rotation. This approach involves interviewing a predefined share of households in multiple survey rounds over a specified period, and the degree of overlap between survey rounds is the same and constant over time. The benefits include efficiency in field practices (as contact and a relationship has already been established, and some information does not need to be collected again) and also some improvement in the precision of estimates of change over time. The rotation scheme 2-2-2 (two rounds in – two out – two in) is one commonly used design is generally considered a good compromise solution. Assuming a round is a quarter, this design provides a fifty percent overlap in two consecutive quarters and the same quarters across two consecutive years. At the same time, this approach requires households to be interviewed four times over an 18-month period, which may increase respondent burden and non-response. Other rotation schemes are used in some countries such as being interviewed in several consecutive periods (e.g. 5, 6, or 8) without interruption.

18. Using sample rotation also has other advantages. It facilitates mixed modes of data collection. For instance, CAPI can be used for the first interview, then being followed by CATI, or web interviews in subsequent rounds using contacts retrieved during the initial or previous interviews. Sample rotation also facilitates the implementation of ad-hoc modules added to the core questionnaire in specific years, but administered only to a sub-sample of households, helping to reduce burden on respondents and interviewers while also controlling costs. A further advantage is the possibility for longitudinal analysis of changes at the individual level, enabling estimates of total transitions or flows – in addition to changes in stock/volume measures. What such analysis can demonstrate is that while stock estimates will change by a certain amount over time, the change is in fact a result of a greater number of movements into and out of each status.

16.2.3. Weighting

19. Weighting is a critical sub-process in labour force surveys. This section covers weighting issues specific to labour force surveys. See Chapter 8 for general guidance about weighting. Weighting enables the dissemination of key labour market aggregates that align with known figures about the target population and its relevant sub-groups, as derived from independent data sources like recent censuses, administrative demographic data, or population projections. Advanced weighting methods and effective strategies can significantly enhance the accuracy of labour force survey (LFS) statistics, as well as their coherence and comparability over time and across different estimation domains and sub-groups. These approaches, in turn, improve the overall quality of all estimates, ensuring consistency with demographic statistics and other surveys, thereby promoting the credibility of National Statistical Offices (NSOs) in published statistics.

20. Given the various imperfections that can affect the actual sample – such as unequal coverage of certain population sub-groups of interest, total non-response, etc. – using design weights to produce direct estimates is generally insufficient for official statistics. As described the Weighting chapter of this Handbook (Chapter 8), design weights typically need to be adjusted for differential non-response affecting sub-groups and to ensure the weighted sample distribution for key variables of interest conforms to known population distributions[4]. In repeated and regular LFS, these adjustments help to **eliminate fluctuations in**

[4] "Designing Household Survey Samples: Practical Guidelines". Studies in Methods Series F No. 98. UNDESA 2008.

population estimates over time due to the sampling variability, thus facilitating the analysis of results. Moreover, they generally help to produce estimates that are **less biased.**

21. When complex sample designs with two or more stages are used, design weights need to take into account the probabilities of selection of the sample units at all stages. Similarly, when sample rotation is used, design weights need to take into account possible differences in the probabilities of selection of the sample units within different rotation groups.

22. Post-stratification and calibration are the most popular methods for complex weighting adjustments; however, calibration can incorporate a much larger volume of auxiliary information than post-stratification (see more details in Chapter 8).

23. Calibration is particularly useful for the LFS **when the available auxiliary information (population figures) consists of multiple components or layers:** a) regions, districts, provinces, urban/rural localities; b) migrants/non-migrants; c) persons, households.

24. Typically, in countries with sub-national geographical domains of estimation, calibration can incorporate different population benchmarks for the different geographical layers. For example, it might include national benchmarks by sex and five-year age groups, and simultaneously, population benchmarks for each sub-national domain by sex and 10-year age groups (or larger age groups, depending on the number of domains, their population size, and the related sample size). This approach will help to produce more accurate labour market indicators for different geographical domains, while improving the quality of the estimates at the total country level.

25. Additionally, population benchmarks can be included for urban and rural localities, either at the national level or within each sub-national domain, with various levels of disaggregation by sex and age groups. Using population benchmarks will help to produce more accurate labour market indicators by type of locality.

26. In countries where the migrant population constitutes a significant portion of the labour force survey's target population, further population benchmarks can be added for migrant sub-groups at the national level (e.g., males and females by a small number of age groups) or at the sub-national level (e.g., total males and total females only), to improve the quality of estimates for these sub-groups.

27. Calibration is also extremely useful to weight LFS sub-samples that have attached a specific sub-module (see Eurostat's methodology for the ad hoc modules and the wave approach in the EU-LFS). For example, in case the ad-hoc module is collected on a sub-sample of each of the four quarters of a specific year, calibration can be used to reach consistency between the estimates from the full sample and from the sub-sample, i.e. to make the annual sub-sample produce the same population distribution and the same key labour market estimates of the annual full-sample.

28. Weighting or calibration benchmarks will be required at the same frequency of the LFS – as such, good collaboration is required between those implementing the LFS and those generating the benchmark population estimates to enable meaningful projections to be developed.

16.3. Use of administrative data to support LFS

29. Administrative data sources cannot replace labour force surveys given LFS' targeted design to measure labour market issues of interest, but the combined use of administrative and survey data can be highly beneficial for labour statistics. In general, administrative records are a rather inexpensive source of labour statistics, crucial for topics not easily covered by traditional statistical sources like occupational

safety and health, industrial relations, and social protection coverage. However, administrative data sources present several limitations, not least the fact that their contents, methodology, coverage, data accuracy, completeness, and timeliness are all determined by administrative processes and needs. Still, the increasing use of electronic methods for data entry and storage is enhancing the availability, timeliness, granularity, and user-friendliness of administrative records, including of those relevant for labour statistics such as population registers, business registers, social insurance records, employment office records, taxation records, records of workers' organizations, collective bargaining agreements, and labour inspection records.

30. In the absence of a regular or recent labour force survey, administrative records can provide basic labour market indicators to monitor trends and patterns. Where a regular labour force survey does exist, administrative data can enhance survey data quality, range, and cost-efficiency. Indeed, administrative sources can complement surveys, providing data on key variables which boast high levels of accuracy and reliability in administrative registers (such as sociodemographic variables) and allowing for shorter survey questionnaires. Administrative registers can also serve as sample frames for surveys, and they can inform the editing, imputation, and estimation processes of a survey. The document presented to the 21st International Conference of Labour Statisticians Making full use of administrative data - A case for administrative registers as a complementary source of labour statistics summarizes the main potential uses of administrative data in the field of labour statistics, citing relevant country practices across regions. The ILOSTAT page devoted to the topic also presents various guides on specific topics of labour statistics for which administrative sources are relevant.

16.4. Managing communication challenges when introducing the latest ICLS standards

31. The implementation of the latest standards via the LFS (or relevant data collection activity) will improve knowledge on labour market issues, but in so doing it can also cause breaks in series in key indicators, whose size will depend on the labour market characteristics (e.g. the prevalence of own-use production work such as subsistence foodstuff production). The communication and dissemination strategy is crucial to achieve the information gains intended by the standards and avoid misinterpretation of the breaks.

32. This strategy should not come as an afterthought: it must be well planned, fostering capacity building from the get-go among key partners, data users and stakeholders. It should promote transparency and ease of understanding, not only by accompanying the released statistics with all the relevant methodological information, but also by selecting indicators and data visualizations clear enough to make the methodological changes evident. One key aspect of the communications strategy must be explaining the benefits of the change in approaches so that users are convinced the change was worthwhile.

33. The note *Communicating new labour statistics standards implementation* provides tips for an effective communication strategy, while the *Quick guide on the communication of results after implementation of latest standards on statistics of work, employment, and labour underutilization* gives practical data visualization examples. A combination of indicators on headcounts of the working-age population, key rates in relation to it (including on participation in unpaid forms of work), key rates in relation to employment, and headline labour underutilization measures, will provide a plain and comprehensive understanding of the labour market.

34. During any transition from one set of standards or methodologies to another, while methodological changes are still new, it will be beneficial to disseminate (with proper explanation) two sets of estimates for selected key indicators: one based on the previous standards, and one based on the new standards. This is

particularly relevant for the transition from the 13th ICLS definitions to those from the 19th ICLS, often generating lower estimates of employment and higher estimates of unemployment (although this is subject to the existing methodologies before the new standards were introduced). Statistical methods such as back-calculation can be used for this purpose. The ILOSTAT page dedicated to the communication of statistics provides further guidance and practical examples to convey clearly and intuitively the impact of methodological changes on key labour indicators.

CHAPTER 17
COUNTRIES UNDER CONFLICT AND POST-CONFLICT

This is a placeholder page.

Chapter 17 will be completed in Spring 2025.

This page has been intentionally left blank.

CHAPTER 18
SMALL ISLAND DEVELOPING STATES

This is a placeholder page.

Chapter 18 will be completed in Spring 2025.