

## **“Relevance, Integrity, Innovation: Are We Measuring Up?”**

Kenneth Prewitt  
Columbia University

The aspirations of statistical agencies listed in the title of this High Level Forum – relevance, integrity, and innovation – confront challenges along many dimensions, of which the most familiar are political and financial. I say a word or two about those challenges, and then turn to a more complicated issue confronting, especially, statistical innovation.

Relevance. Ironically, political pressure on statistical independence, though always troubling and to be resisted, is a signal that what we provide to our countries is relevant and has consequences. If we want our statistics to be increasingly consequential in the policy process or for public understanding, we have to expect greater political attention and the pressure that accompanies it. Stated simply: the more important are our statistics in our respective political economies, the more temptation we put before our political bosses to see to it that those statistics serve one or another political interests. Relevance is in this sense a doubled-edged sword.

Integrity. There is also a double-edge to the goal of statistical integrity. We know that there is no integrity without quality, and that quality is expensive. We know that when governments face difficult budget trade-offs they find it politically easier to under fund statistics than national security, education or health programs. Meanwhile the costs of collecting quality statistics are increasing. As levels of cooperation drop, an even greater effort is needed to maintain quality. A common difficulty facing statistical agencies is whether to allow put quality at risk across a portfolio of surveys (maybe by reducing sample size), or to drop a major program from the agency’s portfolio – and face the political flak generated. In tight budget times, making the choices that protect statistical integrity requires a healthy alliance between a statistical agency and its constituencies so that scientific rather than political criteria will prevail.

Innovation. Innovation has its political and its financial dimensions, but I suggest that we should consider innovation more from the perspective what and how we measure. What we measure – population dynamics, economic activity, household conditions – are not what they were a few decades ago, and will not in decades ahead look as they do now. Issues of classification, standardization, sampling design have to be constantly reassessed. In the U.S., for example, the traditional questions asked about race and ethnicity do not match well with the demographic transformation associated with the huge immigrant flows of the last several decades. Or, again a U.S. example though I suspect relevant in many other countries, the household as the unit of data collection and data reporting is no longer as robust as it once was – here having in mind the many

“households” made up of unrelated people who share living quarters, and move in and out with regularity, or the fact that our (mobile) phone # and our (email) address is no longer tethered to a fixed residence.

As we consider innovations to keep up with the changing population conditions, as well as the political and financial challenges briefly noted, we are likely to confront a deeper set of challenges associated with a change in how we get data about our populations. I can best make my point by contrasting a nation’s statistical *survey* system with its broader *information* system. The latter is constructed from two major data sources: census and survey data, on the one hand, and administrative data, on the other.

In any national information system there is a ratio of data from these two sources. My working hypothesis is that irrespective of what that ratio is (which ranges widely across the members of the UN), it is generally shifting toward a greater reliance on administrative data. That is, if a national government’s information system is today 80% survey based and 20% administrative data based, that ratio is much more likely to be moving toward 60-40 than toward 90-10. As we know, the Nordic countries are already at the far end of the scale, having stopped doing a population census altogether in favor of national registers on which they create a “virtual census” and from which they construct valuable statistical products solely from administrative data.

The U.S., the only country I know well, remains wedded to the census and household surveys as its predominant source of national statistics, but there is a detectable drift toward greater use of administrative data. I believe this change will accelerate. One reason of course is the steady cost increase of surveys. The full cycle cost of the 2010 census in the U.S. will exceed \$13b, double what it was only a decade ago.

In this context, administrative data, a by-product program administration, are comparatively inexpensive, or at least thought to be and the pressure to use them for statistical purposes is growing.

There are a number of serious technical challenges associated with this “innovation,” and here I list just a few from the U. S. context (other countries are well ahead of the U.S. in meeting these challenges).

*Data quality.* Since the 1930s, the U.S. has had a large academic community focused on the quality of surveys -- improving sampling theory, designing procedures for handling missing data, studying questionnaire wording and standardization, etc. There are hundreds of courses on survey methodology in our leading universities, and a number of high quality training programs. At the annual meeting of the American Statistical Association (17,000 members) there are hundreds of papers dedicated to improving survey statistics.

Nothing remotely like this exists for administrative data. It is where surveys were a half-century ago. There is no sustained attention to the error structure of different data sets, no cumulative science of data quality, no dedicated training programs, and only a

scattering of conferences. At least in this U.S., we are decades away from a robust national academic infrastructure for improving the quality of administrative data.

*Personnel.* Paul Chung has spoken eloquently on this question, noting that statistical agencies are going to need persons trained not just in survey methodology, but also in data management. Bringing these two skill sets together will require major readjustment in how statistical agencies organize themselves.

*Data Control.* Administrative data are controlled by the government agency that has designed the data around program management. Arrangements under which administrative data are shared with statistical agencies are, at least in the U.S., difficult to design. For example, although we have worked out ways to link census micro-data with tax returns and employment and earnings history, for understandable reasons the responsible agencies (Internal Revenue Service and the Social Security Administration) are wary of any arrangement in which they do not retain final authority over how the data are used, and by whom. There is certainly no presumption in the U.S. that statistical agencies should start to assume responsibility for the quality of administrative data. Statistical agencies are seen as the home of the census and of surveys, not the locus of a national information system.

*Data Usefulness.* Administrative data are case rich and variable poor. They cover large numbers of people – every child enrolled in school; every taxpayer – but include only a few variables about every person in the data set. Surveys (except the census) are case poor but variable rich. They ask a lot of questions to a comparatively few number of people.

Consequently, administrative data are analytically useful only if linked. One agency has information on my earnings history, but nothing on my health condition; another has my medical records, but knows nothing of my earnings history. The data linkage issues are formidable, at least in countries such as the U.S. that do not have a national register.

*Renegotiating Boundaries.* There are difficult boundaries that have to be renegotiated, one within the government and the other between the government and the commercial sector.

The government's survey enterprise rests on the principle that survey answers can never be used against a respondent. We repeat endlessly that survey data are about statistical measures -- averages and ratios – and not about the individual respondent. In principle, except for longitudinal data sets, we can discard individual identifiers when the survey is complete. Administrative data have to retain individual identifiers, because it is the individual to whom a service is provided or whose behavior must be tracked.

In the U.S., we have built a firewall between survey and administrative data. To breach that firewall will cause alarms, especially over issues of privacy and confidentiality.

There is a second boundary, that between government and nongovernmental data sources. As we know, an enormous amount of potentially useful data are being collected in the commercial sector – what we buy, how we entertain ourselves, where and when we travel, who we email to. Each of us has a “digital identity” that parallels and imperfectly mirrors our real identity. In its war against terrorism, the U. S. government has made extensive use of the electronic data collected by phone companies, airlines, credit card vendors, and banks. The data mining operation that extracts information from these data is completely separate from the official statistical system. But in arenas other than national security, such as my economic behavior, why should I presume that the government will not want to make use of the enormous reservoir of commercially collected data, especially if it is orders of magnitude cheaper to acquire than conducting household surveys.

Simultaneously renegotiating the boundary between statistical and administrative data, and between the government and the commercial sector, is a task for which few statistical agencies are prepared.

To conclude with the question before this Forum – *Are We Measuring Up?*

I am less worried about relevance and integrity, issues much on the mind of our statistical agencies, than I am about innovation, and in particular about how prepared statistical agencies are for the innovations necessary to navigate the new world (for many of us) of national information systems increasingly reliant on administrative data and in some instances on data from the commercial sector. The necessary innovations are in the first instance technical, where there are major challenges in fashioning new data quality procedures. They are also institutional, where statistical agencies, long assumed to be concerned primarily with survey data, will have to reorient themselves around less familiar data sources.

My greatest worry is that governments seeking “statistical” information about their populations will by-pass statistical agencies as they turn to those parts of the government that control large administrative data sets.