Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

BENCHMARKING THE PERFORMANCE OF STATISTICAL AGENCIES

Mariann Lemke, Val Plisko, Marilyn Seastrom, and Daniel Kasprzyk[1]

ABSTRACT

This paper provides an initial look at the performance indicators of statistical agencies in the United States to offer a framework for benchmarking performance. As shown by this initial look, agencies have taken extremely different approaches in measuring their performance, both in terms of what they are measuring and in how they are measuring it. Sharing this information is a first step toward potential collaboration in making these measures more robust and comparable across U.S. agencies and elsewhere. The intent is to borrow from the best to develop parameters for benchmarking against similar agencies.

KEY WORDS: Benchmarking, performance measurement

1. BENCHMARKING MODELS

Benchmarking is the process of identifying, sharing, and using knowledge and best practices to improve any given business process. One can think of benchmarking against past performance by your own organization, benchmarking against internal or external standards of performance, or benchmarking against other organizations. For a government statistical agency like the National Center for Education Statistics (NCES) in the United States, one can think about benchmarking at a number of different levels. At the broadest level, we could use established models and criteria to benchmark NCES as an organization. Looking a bit more specifically at the work of a statistical agency, we could also develop a set of criteria for benchmarking against other statistical agencies. And finally, at the most micro-level, we could benchmark the actual data produced at NCES against accepted statistical standards or cutting-edge statistical techniques, describing survey and non-survey errors. This discussion of accuracy of data is generally subsumed in the discussion of benchmarking at the statistical agency level. In this paper, using information obtained from publicly available federal agency performance plans, we explore the ways in which NCES could begin to benchmark its performance against other U.S. statistical organizations or go beyond to examine performance indicators used by similar organizations in other countries.

1.1 Organizational Models

There are a number of established models and criteria that allow companies and other organizations to be evaluated in particular areas. Many of these kinds of models are awards processes, in which organizations, which achieve a certain level of excellence within each of the criteria, are recognized with an award. Examples include the Baldrige, the European Foundation for Quality Management, and the ISO process. These models are not only meant to be used as part of an awards process, but also as part of a self-evaluative benchmarking effort, so that an organization has some parameters within which to compare itself to other organizations.

One of the criticisms of these kinds of models has been that they are too broad -- that is, it is impossible to talk about an organization without more specific reference to what that organization actually does. In fact, for education organizations and healthcare organizations in the United States, somewhat more specific criteria have been developed in the Baldrige model. The federal government has also developed a model, the Presidential Quality Award, based on the Baldrige criteria and aimed specifically at U.S. federal

[1] Mariann Lemke, Val Plisko, Marilyn Seastrom, and Daniel Kasprzyk, National Center for Education Statistics, U.S. Department of Education, 1990 K Street, NW, Washington, DC 20006, USA.

government organizations. This reflects perhaps a larger trend toward applying business and marketing principles to public and non-profit sectors.

1.2 Statistical Agency Models

Although not directly in response to problems or gaps in the models and criteria described above, statistical agencies have indeed been grappling with the issue of quality as it applies to the work of statistical agencies for some time. Some recent efforts have focused on identifying the features of a quality statistical agency from a variety of perspectives, for example those described in: "Managing Data Quality in a Statistical Agency" (Brackstone, 1999) and "Peer review of the Swiss statistical system" (Fellegi and Ryten, 2000). The Encyclopedia of Statistical Sciences (Update Volume 3, 1999) also defines the quality concept for official statistics, and the paper "Performance indicators for national statistical systems: How are we doing?" (de Vries, 1998) proposes a systematic approach to evaluating the performance of national statistical systems, starting from the Fundamental Principles of Official Statistics, which were adopted by the United Nations some time ago. Each of these efforts has a number of overlapping categories for describing quality, such as timeliness, relevance, etc.

In the United States, the Interagency Council on Statistical Policy (ICSP) established a cross-agency team on performance measurement and reporting in 1999 to review performance plans and recommend common approaches. The group agreed to guidelines for a common approach to reporting performance, in order to facilitate communication and benchmarking among the U.S.'s decentralized statistical agencies. The guidelines encompass "activities of key interest to statistical agencies" and focus on measuring and monitoring the quality of statistical products ("making sure agencies are doing things right") and on working with customers and stakeholders ("making sure agencies are doing the right things"). Some indicators focus on internal measures; others depend on external evaluations. Not all the indicators described are quantitative; some focus on narrative or qualitative descriptions of performance. They include:

- **Timeliness**: to be measured using time-to-release or scheduled release date
- **Output:** numbers of products, product refinement activities, and accessibility of products
- **Accuracy**: to be measured using "as many aspects of data accuracy as feasible" moving toward consistency in how they are measured (review process, sampling and non-sampling error -- can be survey specific)
- **Relevance**: to be measured through descriptive professional reviews or user reviews
- **Reviews, advice, and outreach activities**: to be measured using descriptions of advice solicited from users; also indicators of the extent of outreach activities
- **Customer attitudes**: to be measured through customer surveys (could include customer service and product quality)
- **Use of products and services**: measures could include web statistics and citations, numbers of users, counts of customers; new measures encouraged
- **Cost (financial and response burden)**: total estimated hours for all respondents, as reported to the Office of Management and Budget; cost measures not yet feasible in most cases
- **Contribution to mission**: to be measured through narratives of how data used in the achievement of organization goals

Although the indicators are generally discrete, again, it should be kept in mind that there are no absolute distinctions. Indicators for reviews or relevance, for example, could easily overlap with indicators of customer attitudes.

The process of benchmarking implies that you must be able to determine that another organization performs better in some area than yours does. As an initial step toward benchmarking, the National Center for Education Statistics examined performance measures for a number of different U.S. national statistical agencies. As part of the Government Performance and Results Act of 1994, the U.S. Congress required federal government agencies to begin providing detailed strategic and performance plans for their programs. Using information from these congressionally-mandated federal agency performance plans (FY2000 actual performance plans[2]), we review some of the common areas in which federal statistical agencies might want to benchmark against one another, such as timeliness, accessibility of data, and accuracy of data. Our findings show that agencies have taken extremely different tacks in measuring their performance, both in terms of what they are measuring and in how they are measuring it. These differences point out a number of possible fruitful areas for discussion between agencies and also the difficulties inherent in benchmarking in the statistical area, particularly among decentralized agencies.

A first step was to choose several agencies with similar missions (and with readily accessible information), and see how many of these areas they appear to cover in their performance plans (see Table 1).

Table 1. Performance Indicators for Selected Federal Statistical Agencies[3]

	Timeliness	Output	Accuracy	Relevance	Reviews	Customer Attitudes	Use	Cost	Mission
NCES	x	X	X	x		x			x
NCHS	x	X	X				x		x
BLS	x	X	X		x	x	x		x
BTS	x	X	X	x		x			x
BEA	x					x			x
Census	x	X	X	x	x				x

NCES: National Center for Education Statistics; NCHS: National Center for Health Statistics; BLS: Bureau of Labor Statistics; BTS: Bureau of Transportation Statistics; BEA: Bureau of Economic Analysis; Census: Bureau of the Census.

A second step was to examine a few of these areas in detail to understand exactly how and what agencies are measuring in order that we might begin a true benchmarking process. To this end, here we examine a few of the above-described indicators to explore the variety and depth of the information available for benchmarking purposes.

We decided here to focus on a few key areas in which most of the selected agencies had performance indicators, and those in which indicators were largely quantitative in nature so that they might be easily compared. The areas chosen were timeliness, output (products and accessibility), and accuracy. Some indicators cut across several of the areas selected and are repeated since they are fairly general in nature.

1.3 Timeliness

A look at the indicators shows the quite different approaches taken to measurement of timeliness. In NCES's cases, the main measure of timeliness is taken from the customer's perspective. Although some time-to-release information is provided, it is selective and does not apply for the entire set of data produced by the agency.

[2] Some agencies also make publicly available sub-agency performance or strategic plans; however, information presented here comes from department-level performance plans (indicators for FY02, performance for FY00) available publicly on the Internet.

[3] As noted, indicator categories may overlap such that it is possible to interpret them in several different ways -- the table presents one interpretation of agency performance indicators.

Table 2. Timeliness Indicators for Selected Federal Statistical Agencies

Agency	Indicator	Performance
NCES	At least 85% of surveyed customers in 1999 and 90% in 2001 will agree that the NCES data (publications and data files) are timely, relevant, and comprehensive.	Publications: 72% in 1997; 77% in 1999, 2001 not yet available Data Files: 52% in 1997; 67% in 1999, 2001 not yet available NCES Services: 89% in 1997; 93% in 1999, 2001 not yet available Performance plan also includes narrative information on release times for selected reports.
NCHS	Reduce time lags for release of core data systems by 5% per year. *Vital Statistics* 1) Release of 2000 final mortality data in 18 months (30% reduction from FY96 baseline of 26 months) 2) Release of 2000 final natality data in 16 months (11% reduction from FY96 baseline of 18 months) 3) Preliminary VS 2000 data available within 9 months (10% reduction from FY96 baseline of 10 months) *Health Care Surveys* 4) Release of 2000 National Hospital Discharge Survey data in 18 months (14% reduction from baseline) *Health Interview Surveys* 5) Release of 2000 National Health Interview Survey data in 20 months (23% reduction from baseline)	Indicator for FY00 is different than the indicator for FY02. FY00 target was only for Vital Statistics surveys (1, 2, and 3 at left) and called for reduction in time lag of 2 months (from 21 months to 19 months). This target was achieved.
BLS	Produce and disseminate timely, accurate, and relevant economic information (using percentage of releases that are prepared on time -- release dates against release schedule)	All on time except Employment Cost Index
BTS	Percent of customers satisfied with customer service provided by Department of Transportation.	Information not available
BEA	1) BEA is first in international rankings for production of Gross Domestic Product data 2) Percent of scheduled releases issued on time (based on public schedule of releases) 3) Customer satisfaction is rated greater than 4 (5 point scale)	1) First 2) 100% 3) 4.3 on 5 point scale
Census	1) Percentage reduction from time of data collection to data release (maintain 9% decrease from baseline) 2) Percentage of principal economic indicators released as scheduled (100% target) 3) Disseminate Census 2000 products as scheduled 4) Release 2001 data from Long-Form Transitional Database	1) Information not yet available 2) Information not yet available 3) Information not yet available 4) Information not yet available

NCHS, on the other hand, uses a general goal of reducing time lags for its core data systems. Although year-to-year information is available in the report, performance is reported based on difference from the original baseline (FY96), which implies a focus on overall improvement rather than a continuous process.

BLS uses adherence to an established schedule as its measure of timeliness, but without more information on how schedules are developed, this indicator is not especially useful for making comparisons between agencies. Time-to-release or schedules may also be affected by the kinds of publications an agency produces; for example, some may be legally required to be published at a particular time, while others may be at an agency's discretion.

The overall DOT performance plan does not identify any specific measures for timeliness in statistical reporting, although BTS's strategic plan (also available online) does provide further information (e.g. "by 2003 transportation safety data will be available at least monthly, with no more than 30 days lag time"). DOT's indicator of customer satisfaction could encompass customer satisfaction with the timeliness of information provision (as is the case with NCES), but this measure is not specified in DOT's plan.

BEA's three indicators for timeliness may be less helpful for benchmarking, since one relates specifically to a process for ranking countries on their release of GDP data, although similar measures could potentially be created for the data produced by other agencies. The percent of scheduled releases issued on time, like the BLS indicator, could be useful for benchmarking if other agencies had similar measures with additional information provided. Finally, the customer satisfaction measure may address timeliness, although it is not clear in the indicator.

Census's indicators on percentage reduction from time of data collection to data release, like BEA's, could be a useful comparative measure. For this kind of indicator to truly be useful, agencies might need to compare the frequency and complexity of surveys conducted in order to develop a schedule that reflected the parameters of data collection and analysis. For example, one-time surveys may take longer to produce data than those which are conducted frequently, for which systems are in place to process and analyze data quickly. Another factor to be considered might be whether data releases are preliminary or final. Percentage of principal economic indicators released as scheduled (100% target) and release of Census and LFTDB information indicators are similar to BLS and BEA indicators described above.

1.4 Output (Products and Accessibility)

As with the indicators of timeliness, NCES's indicators on output focus on external customer reviews of products, both data files and publications. This is related to but different than the measures described for the indicator in the ICSP guidelines, which focus on numbers of products, accessibility, and efforts to improve products.

NCHS, on the other hand, has a host of indicators to address the need to make more products available and more easily accessible, as well as to improve those already in existence. Many of these indicators are fairly specific to NCHS, but accessibility targets, for example, could be useful benchmarks for other agencies.

BLS uses a simple indicator of Internet site user sessions to give some measure of the accessibility and use of their products. It is clear that this type of indicator must be interpreted carefully, since as the ICSP guidelines point out, numbers of user sessions can reflect the particular design of an agency's website, or a number of other factors which are not directly related to the true use or accessibility of materials.

Table 3. Output Indicators for Selected Federal Statistical Agencies

Agency	Indicator/Target	Performance
NCES	1) At least 85% of surveyed customers in 1999 and 90% in 2001 will agree that the NCES data (publications and data files) are timely, relevant, and comprehensive and are of high quality in terms of accuracy, reliability, and validity. 2) At least 85% of surveyed customers in FY1999 and 90% in 2002 will agree that NCES publications are easy to read. 3) At least 85% of surveyed customers in FY1999 and 90% in FY2002 will rate NCES publications as useful to their work.	1) Overall quality: Publications: 90% in 1997; 93% in 1999, 2001 not yet available; Data files: 87% in 1999, 2001 not yet available 2) Clarity of writing: 87% in 1997; 90% in 1999; Useful to work: 86% in 1997; 89% in 1999; Overall quality: 90% in 1997; 93% in 1999 3) 86% in 1997; 89% in 1999
NCHS	1) Develop new monitoring tools needed to address emerging topics (children with special health care needs, NHANES) 2) Make health statistics available via the Internet (monthly vital statistics reports will be available to be viewed, searched, and downloaded via the Internet within 4 months of release) 3) Release statistics in new formats to speed the release of data on high-priority topics (one report) 4) Allow non-NCHS researchers to access detailed data files in a secure environment, without jeopardizing the confidentiality of respondents -- through the NCHS Data Center (establish Data Center; at least 30 researchers to use) 5) Increase the number of articles published in peer-review journals that utilize NCHS data (increase by non-NCHS researchers 10% over 5 years; increase by NCHS 10% over 5 years) 6) Increase the number of people who obtain key health information from the NCHS website (microdata users increase by 5%) 7) Increase the number of subgroups with available data in the Health People 2010 template	1) achieved; 2) achieved; 3) achieved; 4) Approximately 2/3 achieved; 5) Information not yet available; 6) Information not yet available; 7) Information not yet available
BLS	Produce and disseminate timely, accurate, and relevant economic information (using average number of Internet site user sessions per month)	Performance not indicated
BTS	Percent of customers satisfied with customer service provided by Department of Transportation (DOT)	Not available for either DOT or BTS indicators
BEA	Customer satisfaction is rated greater than 4 (5 point scale)	4.3 on 5 point scale (Indicator on website utilization discontinued.)
Census	Indicator on qualitative customer evaluations discontinued.	

Again, BTS's strategic plan identifies several indicators and targets related to output, although the overall DOT performance plan does not identify any such measures. DOT’s measure of customer satisfaction, as before, could encompass customer satisfaction with output but this is not specified in DOT’s plan. The output indicators BTS provides are fairly general in nature, although some of them could provide a means for comparison. For example, one goal describes making all major data sets accessible through the web. With some refinement, that measure could provide interesting comparative data if agencies identified their major data sets and how to make them accessible through the web.

BEA's output measure does not directly address output, but like NCES's focuses on customer satisfaction, although it is not clear what role products and accessibility play in the customer satisfaction surveys. Interestingly, BEA's plan notes that an indicator on website utilization was dropped, because "website usage and technology are growing and changing so rapidly that it is questionable how meaningful any measures and targets can be."

Census has no performance indicators that relate directly to output. A measure of customer satisfaction, which may have had some indirect relationship to output, was dropped in order to include some specific indicators related to the release of Census 2000 information (see timeliness indicators above).

1.5 Accuracy

Table 4. **Accuracy Indicators for Selected Federal Statistical Agencies**

Agency	Indicator/Target	Performance
NCES	1) At least 85% of surveyed customers in 1999 and 90% in 2001 will agree that the NCES data (publications and data files) are timely, relevant, and comprehensive and are of high quality in terms of accuracy, reliability, and validity.	1) Accuracy: Publications: 84% in 1999; Data Files: 74% in 1997; 82% in 1999
NCHS	No accuracy indicators published	
BLS	1) Produce and disseminate timely, accurate, and relevant economic information (using measures of quality for principal economic indicators: National Labor Force, Employment, Hours, and Earnings, Consumer Prices and Price Indexes, Producer Prices and Price Indexes) Quality measures specific to each survey, e.g., *Employment, Hours, and Earnings*: Root mean square error of total nonfarm employment (a measure of the amount of revision) <70,000. (Baseline is FY 2000.) *Producer Prices and Price Indexes*: (1) Percent of domestic output, within the scope of the PPI, that is covered by the PPI: goods produced = 85.1 percent; services produced = 38.8 percent; total production = 52.6 percent. (Baseline is FY 1997.) (2) Percent of months that the change in the one-month Finished Goods Index (not seasonally adjusted) between the first-published and final release was ±0.2 percent. (Baseline will be set in FY 2001.) 2) Improve the quality of the CPI Index and the PPI Index (improve CPI sample design and estimation methodology….) 3) Improve the statistical quality of the output from the CES program	1) All quality measures achieved except for Producer Price Index 2) Information not available 3) Information not available
BTS	No accuracy indicators published	
BEA	Indicator on data accuracy (score on a scale of 100) was discontinued for FY2002 due to "lack of reliability in the measure."	
Census	1) Percentage of household surveys attaining specified reliability measurements 2) Percentage of household surveys with initial response rates > 90% 3) Percentage completion of housing unit address list (comparison of Accuracy and Coverage Evaluation with Bureau's Master Address File)	1) 100% 2) 100% 3) Information not yet available

NCES again measures accuracy using external customer measures related to both its data files and its publications. A few agencies show no indicators for accuracy, although again, interestingly, BEA discontinued an indicator on data accuracy (a score on a scale of 100) due to problems with the measure. This reflects the difficulty of developing such comprehensive measures for an entire agency or set of data collections.

BLS, on the other hand, has specific accuracy measures for a number of its key surveys, although it appears that these measures may relate to agency goals for the economy rather than for quality in statistical processes. These measures, while not necessarily comparable across agencies, do provide some indication of the level of accuracy for which BLS strives, and potentially similar kinds of measures could be developed for other agencies.

BTS, as in previous examples, has a set of general performance goals in its own plan that relate to data standards and practice but none are included in DOT's performance plan. BTS's case raises an interesting point about what role statistical agencies play within their own departments and how important their information is to defining and reporting on a department's mission.

Census's measures of accuracy, as with BLS's are largely targeted to specific surveys. But for agencies that conduct similar household surveys, the performance indicator specified (response rates over 90%) could be something useful for agencies to compare.

Agencies may also adopt an inward or self-evaluation focus on the quality of their data--evaluating and benchmarking itself against a set of statistical standards or guidelines accepted by the agency. These guidelines codify professional standards against which the agency can measure its performance. Such guidelines are not uniformly available across all statistical agencies, but a few examples exist, such as the standards developed by the U.S. Department of Energy (1992) and the U.S. Department of Education (1992). A comparison of the current practices of individual data collection programs against these standards can provide a useful benchmarking tool for an agency, but in general such comparisons do not form part of departmental performance plans. Indicators related to these data quality standards can provide another avenue for agencies to measure performance. Another internal performance measure that can be used to assess the performance of a survey program is the survey quality profile. A quality profile is a report that summarizes what is known about the sources and magnitudes of errors in a survey. It is a major summarization of information about the quality of data from a survey program. As a source of documentation, a quality profile reports the "what, how, and why" of survey and statistical procedures and the errors associated with various aspects of the survey. Kasprzyk and Kalton (2001) suggest quality profiles that assess the multidimensional aspect of data quality and bring together relevant qualitative and quantitative information that can be used to assess the performance of data programs.

Further, an Office of Management and Budget Subcommittee has conducted some comparative work in the reporting of information about the quality of survey data, and that this information is not reflected in the performance indicators noted above. For example, Kasprzyk, McMillen, Atkinson, Giesbrecht, Schwanz, and Seiber (1999) and the Federal Committee on Statistical Methodology (2001) describe various error sources in surveys and how and whether agencies report sources of error in printed reports and Internet sites, finding considerable variability in agency practices in reporting this information.

2. NEXT STEPS IN BENCHMARKING

As the American Productivity and Quality Center describes it, "Informally, benchmarking could be defined as the practice of being humble enough to admit that someone else is better at something and wise enough to try to learn how to match, and even surpass, them at it."

For NCES, next steps may include re-visiting our own measures of performance. As evidenced in the tables, NCES shows almost total reliance on a customer survey for information about how the agency is performing. While this kind of information is valuable, it certainly does not tell the whole story. Looking at the performance indicators from a variety of agencies may provide valuable input into new ways of reporting performance.

The next step at a federal level will be for agencies to jointly examine specific performance measures and how they could be made more robust and more comparable. This could include potential collaboration beyond U.S. borders with similar international statistical agencies. Agencies have done some real comparative work in very specific areas, such as ways in which survey error is reported, but little has been done to examine how performance in more general areas could be compared. No two U.S. statistical agencies are collecting the same kind of information; on the other hand, it is clear that there is some consensus around the kinds of attributes that make for a quality statistical organization, no matter what kind of statistics it collects. Indeed, one first step could be to make information on how statistical agencies measure performance readily available, taking the work of the ICSP even further.

Of course, comparison is only the first step in benchmarking. A true benchmarking process requires an agency not only to explore differences but also to apply observed best practices that have led to differences. This kind of change, of course, can be difficult to make in any organization. But we hope that these first steps may lead to further discussion in the U.S. and elsewhere on how to measure performance in statistical agencies so that the true work of benchmarking can begin.

REFERENCES

Brackstone, G. (1999). "Managing Data Quality in a Statistical Agency", Canada: Statistics Canada, Survey Methodology, Catalogue No. 12-001-SPB, Vol. 25, pp. 2-21.

de Vries, W. (1998). "Performance Indicators For National Statistical Systems: How are We Doing?", Netherlands Official Statistics, Volume 13, pp. 5-12.

Encyclopedia of Statistical Sciences. Update Volume 3 (1999). New York: John Wiley and Sons.

Fellegi, I. and Ryten, J. (2000). *A Peer Review of the Swiss Statistical System.* Neuchatel, Switzerland: Swiss Federal Statistics Office.

Interagency Council on Statistical Policy. (2000). "Guidelines for Reporting Performance by Statistical Agencies", unpublished report, Washington, DC.

Kasprzyk, D., D. Atkinson, L. Giesbrecht, M. McMillen, D. Schwanz, W.K. Seiber. (2000). "Reporting Sources of Error: The United States Experience". *Proceedings of the 52nd Session of the International Statistical Institute,* pp. 19-30.

U.S. Department of Commerce. (2001). *Commerce's 2000 Annual Program Performance Report and FY2002 Annual Performance Plan.* Washington, DC.

U.S. Department of Education. (2001). *U.S. Department of Education's 2000 Performance Report and 2002 Program Annual Plan.* Washington, DC.

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. (2001). *HHS Operating Division Performance Plans and Reports.* Washington, DC.

U.S. Department of Labor. (2001). *FY 2002 Annual Performance Plan*. Washington, DC.

U.S. Department of Labor, Bureau of Labor Statistics. (n.d.). *BLS Strategic Plan.* Washington, DC.

U.S. Department of Transportation, Bureau of Transportation Statistics. (n.d.). *BTS Strategic Plan.* Washington, DC.

U.S. Department of Transportation. (2001). *Performance Report Fiscal Year 2000, Performance Plan Fiscal Year 2002*. Washington, DC.