



REPORT

FINAL REPORT

Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines

April 27, 2018

John L. Czajka
Mathew Stange

Submitted to:

Internal Revenue Service
Statistics of Income Division
P.O. Box 2608
Washington, DC 20013-2608
Project Officer: Brian Balkovic
Contract Number: TIRNO=13-Z-00011-0006

Submitted by:

Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: John L. Czajka
Reference Number: 50497

This page has been left blank for double-sided copying.

CONTENTS

EXECUTIVE SUMMARY	ix
I. INTRODUCTION.....	1
A. A new paradigm?.....	1
B. International quality standards and guidelines	7
C. Literature review process	7
D. Organization of this report	8
II. INTERNATIONAL STANDARDS	13
A. European Union.....	14
1. <i>European Statistics Code of Practice</i>	14
2. <i>ESS Quality Assurance Framework</i>	16
3. <i>ESS Handbook for Quality Reports</i>	17
B. Selected national statistical organizations.....	35
1. Canada.....	35
2. Australia	40
3. United Kingdom.....	42
4. The Netherlands.....	46
5. Finland	50
6. Sweden	53
7. OECD.....	55
8. IMF.....	56
III. EXTENDING TOTAL SURVEY ERROR TO INTEGRATED DATA.....	59
A. Zhang's two-phase framework	59
1. Total Survey Error	59
2. Zhang's framework.....	61
B. Stats NZ's use of the two-phase framework.....	67
IV. QUALITY ASSESSMENT IN THE USE OF ADMINISTRATIVE DATA FOR OFFICIAL STATISTICS	77
A. The UNECE	77
B. The European Commission	79
C. UK Statistics Authority	83
D. Statistics Netherlands' quality framework for administrative data	85

V.	BIG DATA AND OFFICIAL STATISTICS.....	89
	A. UN Global Working Group on Big Data.....	89
	B. IMF Internal Group on Big Data.....	96
	C. AAPOR Big Data Task Force	97
VI.	DISCUSSION AND CONCLUSIONS.....	99
	The European context.....	99
	The quality concept.....	99
	Standards for integrated data	100
	Issues in quality measurement for integrated data	103
	Quality and Big Data	106
	Quality reporting.....	107
	REFERENCES.....	109

TABLES

I.1.	Principal sources by country/international organization and type of standards.....	9
II.1.	Quality assurance framework of the European Statistical System	15
II.1.	Distribution of UK quality measures/indicators by quality dimension and stages of the statistical production process.....	45
III.1.	Stats NZ's quantitative quality indicators for phase one	69
III.2.	Stats NZ's qualitative quality indicators for the phase one	70
III.3.	Stats NZ's qualitative quality indicators for the phase one representation	70
III.4.	Stats NZ's quantitative quality indicators for phase two	71
IV.1.	Quality indicators for administrative data used as an input source	81
V.1.	Dimensional structure of the input phase of the UNECE Big Data quality framework.....	92
V.2.	Dimensional structure of the output phase of the UNECE Big Data quality framework	95

FIGURES

III.1.	Survey life cycle from a quality perspective	60
III.2.	Two-phase life cycle of integrated micro data from a quality perspective	63

EXECUTIVE SUMMARY

The research landscape for federal statistical agencies is moving to a new paradigm in which survey data are no longer the principal data type. This shift is due to growing challenges facing traditional survey research, including an increasing reluctance of people to complete surveys and deteriorating coverage of sample frames. The new paradigm is characterized by the use of administrative data and other forms of Big Data as alternatives to survey data, and, increasingly, the use of integrated data that combines data from multiple sources, such as linked survey and administrative data. This new paradigm necessitates new quality standards that address integrated data. In response, the Interagency Council on Statistical Policy (ICSP), led by the Chief Statistician of the United States of the Office of Management and Budget (OMB), tasked the Federal Committee on Statistical Methodology (FCSM) Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources with preparing analyses and recommendations to inform the development of cross-agency standards suitable for integrated data. The ICSP asked Mathematica Policy Research to review quality standards from national statistical agencies outside the U.S. as well as international organizations like the United Nations (UN). This report reviews information on international standards and guidelines on quality reporting relative to statistical estimates that combine survey data with other types of data. The report is based on a search of both published literature and grey literature from statistical organizations' websites that identified a number of articles, book chapters, reports, and official documents addressing data quality standards generally and with application to administrative data, Big Data, and integrated data specifically.

International statistical agencies and organizations are nearly uniform in defining quality as multi-dimensional, though specific dimensions vary.

Our review of reporting standards begins with Eurostat and the European Statistical System (ESS)--which draw on their member nations' experience in working with integrated data in the form of linked administrative registers, sometimes combined with survey data--and are leaders in the development of reporting standards for statistical data generally. In addition, we review standards documents from select European countries, Canada, Australia, the Organization for Economic Cooperation and Development (OECD), and the International Monetary Fund (IMF)—each of which adds something unique in its perspective on data quality.

The literature across these countries and organizations is nearly uniform in defining data quality as “fitness for use” in which “good” or “high” quality data meets its intended purpose in operations, decision-making, and planning. Across these standards, we found a consensus that quality is multi-dimensional, and its measurement encompasses both

Definitions of Five Quality Dimensions Common to International Statistical Agencies and Organizations

Relevance is the extent to which data can be shown to satisfy user needs.

Accuracy and reliability refer to the degree to which statistical information correctly describes the phenomena it was designed to measure.

Timeliness refers to the length of time between the reference period for a statistical estimate or dataset and when it is made available to users; punctuality refers to whether data were delivered on the date they were scheduled for release.

Coherence and comparability refer to the degree to which statistical information is logically consistent and can be brought together with information from other sources or different time periods.

Accessibility and clarity refer to the simplicity and ease of use of data, including how and under what conditions users can access it and how readily users can correctly interpret statistics in light of the supporting information and other assistance that is provided.

quantitative and qualitative indicators. Five dimensions appear almost universally in quality frameworks around the world: (1) relevance, (2) accuracy and reliability, (3) timeliness and punctuality, (4) coherence and comparability, and (5) accessibility and clarity. There is variation among organizations, however, in terms of combining some of these dimensions and adding in others to their respective frameworks, such as including costs and confidentiality. For example, the UK includes confidentiality—meaning private information about individual persons should be kept confidential and used for statistical purposes only—as part of a quality framework, which is of growing importance as the ability to link administrative data, Big Data, and survey data is providing ever more capability to find out detailed information about individuals.

Quality measures and indicators abound in the official literature of these organizations. Quantitative indicators exist but tend to be limited to the dimension of accuracy and reliability. Qualitative indicators tend to be descriptive in nature, with many requiring a high level of detail. The information requested tends to be at the level of the individual statistic. All things considered, the preparation of a quality report for many of these organizations represents a considerable undertaking, albeit moderated by the fact that much of the material required for recurring estimates can be repeated.

The Total Survey Error (TSE) Framework is a paradigm for looking at all errors stemming from the design, collection, and processing of survey data (Groves et al. 2009). Error refers to differences between the survey response observed and the true value the survey was measuring. The TSE Framework looks at errors related to representation and errors related to measurement.

Representation errors include coverage error (occurs when there is a not a perfect one-on-one match of the target population and sample frame), sampling error (occurs when collecting data from a sample instead of a census of the target population), and nonresponse error (occurs when not all sample members respond to a survey and those that do respond differ on the outcomes of interest from those who did not respond).

Measurement errors include validity or construct error (occurs when a survey question does not measure the underlying concept it is intended to measure), measurement error (occurs when something about the survey instrument, interviewer, or respondents results in survey response that differs from the true value), and processing error (occurs from data entry, coding, or analysis that results in a survey response differing from a true value).

Only Statistics New Zealand has developed an error assessment framework explicitly for integrated data.

Turning from reporting standards generally to integrated data specifically, we found only Statistics New Zealand (Stats NZ) has developed a framework explicitly designed to address integrated data. Quality here is more focused as an error assessment framework for integrated data than the more multidimensional look at quality we saw earlier. Stats NZ’s framework builds off the work by Zhang (2012) at Statistics Norway who proposed a “two-phase life-cycle model of integrated statistical micro data” building on the Total Survey Error model of Groves et al. (2009). Phase one describes a single micro data source—generalized to include both survey and administrative data—through a process of conception, collection, and processing. Each input to the integrated micro data would have its own phase one assessment. Phase two depicts the sources of error characterizing the integrated micro data, where the error components reflect the integration process, which may include transformation of the initial input data. We summarize Zhang’s work and how Stats NZ has used it to create a quality framework, which includes the addition of a third phase focused on the statistics derived from the integrated micro data.

Other literature looks at the use of administrative data in official statistics.

Distinct from the literature on quality frameworks and quality reporting in general is a literature focusing on quality issues in the use of administrative records in the production of official statistics, such as using official tax records to produce estimates relating to small businesses. Within this area, we review literature drawn from the United Nations Economic Commission for Europe (UNECE), the European Commission, the United Kingdom (UK) Statistics Authority, and Statistics Netherlands. Two important points drawn from this review are: 1) an assessment for each of the three stages of input, data processing, and output is essential before using administrative data for statistical purposes, and 2) the availability of good metadata at each stage is vital to such an assessment.

Administrative data are data collected for a non-statistical purpose, such as tax records (UNECE 2011). Administrative data, though, can have multiple statistical uses, including as official benchmarks to estimate the bias and variance of survey data collected on similar outcomes or as data to blend with survey data to supplement data missing from survey data. Administrative data might even replace the need a survey data collection.

The distinction between the original purpose of an administrative data source and its statistical use is discussed repeatedly. An important implication is that in assessing the quality of an administrative data source for use in preparing official statistics, one must evaluate the quality of the data source as it was originally intended to be used as well as how it will be used in the statistical estimate. Coverage emerges as an especially important quality issue when administrative data are used for official statistics.

Work on a quality framework for Big Data is ongoing.

The report includes a review of international efforts to establish the usefulness of Big Data as a source for official statistics. Specifically, we look at the ongoing work of two entities, the UN Global Working Group on Big Data for Official Statistics and the IMF, and we review the Big Data Task Force Report of the American Association for Public Opinion Research (AAPOR). While the UN Working Group has yet to produce a set of official standards, it has made progress in developing a Big Data quality framework revolving around three general principles: 1) “fitness for use” should remain the central focus in assessing the quality of a data source; 2) the framework should be generic and flexible and able to apply its quality dimensions to the three phases of input, throughout, and output; and 3) the framework should allow an assessment of effort versus gain—that is, a determination of whether the effort involved in obtaining and analyzing the data is worth the benefits gained from doing so. The development of the framework is ongoing. The AAPOR Big Data Task Force also notes that, to date, “very little effort has been devoted to enumerating the error sources and the error generating processes for Big Data.” The Task Force concluded that a total error framework is needed for Big Data, and it offered “a skeletal view” of such a framework.

Big Data is defined as data that is large in volume, collected at a rapid velocity, and has a complex variety of formats (AAPOR 2015). Examples of Big Data include social media data, sensor data, and transaction data, among others.

Conclusion

The goal of this review was to compile information on international standards and guidelines on quality reporting relevant to statistical estimates that combine multiple sources of data. We find that:

- Only one national statistical organization—Stats NZ—has developed a quality framework explicitly designed to address integrated data
- Eurostat’s quality standards and guidelines, which apply to most of Europe and are perhaps the most extensive, deal with integrated data to a much more limited degree and instead focus on quality more generally
- Efforts to deal with quality aspects of administrative data are much farther along than efforts to deal with the quality of other forms of Big Data
- Increased granularity (e.g., sufficient data to make substate or other smaller geographic estimates, examine subpopulations, or more precise estimates, among others), which is a benefit of integrated data, and one promoted by the recent Committee on National Statistics panel on multiple data sources, is rarely mentioned in international quality frameworks

Many of the quality assurance frameworks and the associated standards and guidelines reviewed in this report are associated with extensive prescriptions for quality assessments and their communication to data users in detailed quality reports. Of note, the volume and types of information requested in Eurostat quality reports bears substantial resemblance to what was included in the quality profiles prepared by a number of U.S. federal agencies in the 1990s and early 2000s. While quality profiles were intended for recurring surveys, they were updated or repeated for only one survey, and no new profiles have been produced in the past decade. Their preparation demands resources that are increasingly less available, they require detailed information that may not exist, and their value to the survey producer in terms of suggesting future improvements is questionable. This prior experience suggests that federal agencies are not likely to embrace the recommendations of international agencies for substantially more extensive reporting on quality than is done currently. A more acceptable format may be one similar to the Source and Accuracy statements that appear as appendices in some Census Bureau publications.

I. INTRODUCTION

A. A new paradigm?

Official statistics in the U.S.—that is, the statistics produced by federal agencies—have depended heavily on probability surveys of households and other entities as their principal source of data. While the current landscape of data sources also includes administrative records and, to a limited degree as yet, so-called “Big Data,” traditional surveys—including censuses as a special case—continue to dominate published statistics and public use micro data. This is changing. As federal agencies are increasingly looking to administrative records and Big Data for ways to enhance their survey-based products, prominent voices in the federal statistical community are promoting the virtues of combining multiple data sources and even heralding the emergence of a new paradigm for official statistics.

Constance F. Citro, then Director of the Committee on National Statistics (CNSTAT) within the National Academy of Sciences, argued for a transformation of ongoing household survey programs “to use multiple data sources to provide information of greater value” (Citro 2014). In July 2015 the National Academy of Sciences’ National Research Council approved a consensus panel to address issues related to combining data sources. The Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods: Frameworks, Methods, and Assessments released two reports in 2017: *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy* and *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (National Academies of Sciences, Engineering, and Medicine, 2017a and 2017b). Both reports speak to the need for a new paradigm based on combining diverse data sources from government and the private sector to replace the survey paradigm underlying most of federal statistics.

The notion of combining data from administrative and survey sources to create an integrated dataset or integrated estimates, though, is hardly new to the U.S. The pioneering work of Fritz Scheuren and others in linking survey data from the Census Bureau with administrative data from the Internal Revenue Service (IRS) and the Social Security Administration (SSA) in a series of “exact match” studies was conducted in the 1970s (Kilss and Scheuren 1978). Statistical matching of survey data to IRS administrative data—as a way of combining a representative sample of nonfilers with a representative sample of taxpayers—has been used by the Office of Tax Analysis in the Treasury Department to create databases for modeling reforms to the tax system since the 1970s as well, and various ways of incorporating administrative estimates of program participants into survey-based microsimulation models to address survey underreporting of participation have been used since that time also. In addition, the National Income and Product Accounts and the Consumer Price Index have been developed from multiple data sources for decades (Horrigan 2013 cited in Citro 2014). Numerous other examples could be listed.

The recent focus on the development of integrated data for national estimates builds on this history, but it derives more immediately from a growing recognition that the viability of sample surveys as well as censuses conducted using traditional methods is declining. Surveys and censuses are becoming more difficult to carry out, and the quality of key estimates is deteriorating (Citro 2014). Survey organizations face diminishing quality in their sample frames, exemplified by the increasing undercoverage of frames accessed through random digit dialing, which was once ubiquitous. Household members increasingly choose not to respond to surveys, leading to growing nonresponse rates across all types of survey research (De

Statistical agencies commonly define **data quality** as “fitness for use” (Juran and Gryna 1980), meaning that data meet its intended purpose in operations, decision-making, and planning. Throughout the 20th century and the first decade of the 21st, statistical organizations around the world issued guidance on improving and maintaining quality in their operations and assessing and reporting on quality in the statistics they produce. This guidance came in the form of data quality frameworks, guidelines, performance indicators, and reporting standards, among others (Biemer and Lyberg 2003).

Leeuw and De Heer 2002, National Research Council 2013, Brick and Williams 2013). Rising rates of item nonresponse among those who do respond to surveys are resistant to commonly used techniques to increase cooperation, leading to greater reliance on imputation. Efforts to counteract these problems—including ever more call attempts and longer field periods—contribute to what is perhaps the biggest challenge that surveys and censuses must confront: increasing data collection costs. Such efforts may also result in less timely estimates.

Citro (2014) suggests that the various approaches applied to improve the quality of survey data are commendable but insufficient to address the most serious problems. She recommends that statistical agencies first determine their users’ needs and then work backwards to identify the best data sources to “serve those needs in the most cost-effective and least burdensome manner possible.” Citro argues, further, that this “multiple data sources paradigm” should be employed by all statistical programs, regardless of whether they have been based historically on survey data, administrative data, or other sources. Focusing on administrative records, she lists eight ways in which statistical agencies could use such data to improve the quality of household-level survey data. These include replacing erroneous survey responses where administrative data can provide the requested information or eliminating the survey questions for these items entirely and using the values from administrative records directly.

Administrative data are often included under the Big Data umbrella although administrative data have been around for centuries, have been used routinely by government agencies at all levels, and tend to be highly structured—qualities not typically associated with Big Data. While a consensus definition does not exist, the description of Big Data as characterized by three Vs—volume, velocity, and variety has stuck.¹ Other Vs have been added. A group at the International

¹ This characterization is attributed to Laney (2001).

Monetary Fund (IMF) added veracity and volatility, where “veracity refers to the noise and bias in the data as one of the biggest challenges to bringing value and validity to Big Data,” and “volatility refers to changing technology or business environments in which Big Data are produced, which could lead to invalid analyses and results, as well as to fragility in Big Data as a data source” (Hammer et al. 2017). Of these five Vs, only volume is descriptive of administrative data. Nevertheless, a widely cited classification of Big Data by the United Nations Economic Commission for Europe (UNECE) includes administrative data under the category of traditional business systems, or process-mediated data (UNECE 2013). The other two categories are social networks, or human-sourced information, and the Internet of things, or machine-generated data. In this report we address administrative data separately from Big Data, as does nearly all of the literature we reviewed.

In its second and final report, the CNSTAT panel cited a growing need for greater granularity in federal statistics as something that surveys would be hard pressed to deliver, even with the higher response rates of earlier decades. The American Community Survey (ACS) was designed to address this need, but there is a trade-off between timeliness and geographic detail. To provide estimates for substate areas below the very largest, ACS data must be aggregated over five years, which means that the resulting estimates for substate areas are multi-year averages that are not well suited to monitoring short-term trends.² Furthermore, estimates from the five-year aggregates are not available to users until about four years after the mid-point of the data series. By combining survey data with non-survey data sources that provide greater geographic detail and applying appropriate statistical models, it is possible to improve the granularity and the timeliness of the resulting estimates.

² The Census Bureau produced three-year aggregates of ACS estimates until recently, but these did not provide the same level of geographic detail as the five-year estimates.

The CNSTAT panel considered a number of other possible ways—including those cited by Citro (2014)—in which data from multiple sources could be combined to generate estimates that improve upon what can be produced with individual sources alone. Combining data sources requires the application of a variety of statistical techniques. The panel’s second report reviews record linkage, multiple frame methods, imputation-based methods, and modeling techniques such as small area estimation.³ The panel concludes its review with a recommendation:

Recommendation 2-2: To achieve transparency, federal statistical agencies should document the processes used to collect, combine, and analyze data from multiple sources and make that documentation publicly available.

The notion of transparency or openness in the reporting of data quality by federal statistical agencies has a long history. Statistical Policy Working Paper 31, “Measuring and Reporting Sources of Error in Surveys,” which was produced by the Federal Committee on Statistical Methodology (FCSM) and published by the U.S. Office of Management and Budget (OMB) in 2001, cites a 1978 OMB document on the importance of openness:

“To help guard against misunderstanding and misuse of data, full information should be available to users about sources, definitions, and methods used in collecting and compiling statistics, and their limitations” (OMB 1978).

In 2006, OMB issued Statistical Policy Directive No. 2, which delineated 20 standards and associated guidelines for federal censuses and surveys (OMB 2006). The standards and guidelines cover the survey process from design through dissemination. The directive describes the standards as documenting “the professional principles and practices that Federal agencies are required to adhere to and the level of quality and effort expected in all statistical activities.” The guidelines represent “best practices that may be useful in fulfilling the goals of the standard.”

³ Lohr and Raghunathan (2017) provide a more extensive review of statistical methods for combining information from multiple data sources.

Three of the standards deal explicitly with data quality, and all three specify some form of reporting of quality to users:

- **Standard 3.2:** Agencies must appropriately measure, adjust for, report, and analyze unit and item nonresponse to assess their effects on data quality and to inform users. Response rates must be computed using standard formulas to measure the proportion of the eligible sample that is represented by the responding units in each study, as an indicator of potential nonresponse bias.
- **Standard 3.3:** Agencies must add codes to collected data to identify aspects of data quality from the collection (e.g., missing data) in order to allow users to appropriately analyze the data. Codes added to convert information collected as text into a form that permits immediate analysis must use standardized codes, when available, to enhance comparability.
- **Standard 3.5:** Agencies must evaluate the quality of the data and make the evaluation public (through technical notes and documentation included in reports of results or through a separate report) to allow users to interpret results of analysis, and to help designers of recurring surveys focus improvement efforts.

OMB has leverage to enforce many of the standards through its role in reviewing and giving final approval to surveys conducted or sponsored by the federal government.

The principle of openness and a recognition of the importance of standards and guidelines for survey processes underlay the development of Working Paper 31, and it bears directly on the motivation for this report: Moving to the new paradigm will require new quality standards that address integrated data. As such, an FCSM Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources has been tasked by the Interagency Council on Statistical Policy (ICSP) with preparing analyses and recommendations that can inform the development of standards suitable for integrated data. To inform this effort, Mathematica Policy Research was asked by ICSP to review quality standards from national statistical agencies outside the U.S. as well as international organizations.⁴ This report presents our findings.

⁴ The review was produced under a task order issued to Mathematica by the Statistics of Income (SOI) Division of the Internal Revenue Service, one of the 13 federal statistical agencies recognized by OMB. The SOI Division is represented on the FCSM.

B. International quality standards and guidelines

There are three main reasons to focus on international standards. First, administrative data systems are much more developed in many other countries than they are in the U.S. Many of these country's statistical systems include population registers and other types of registers that date back hundreds of years. Registers in the Nordic countries, for example, have a long history of use in official statistics.⁵ Second, the decline in survey response rates has been more rapid elsewhere—especially Europe—than in the U.S., so many countries have had more time to think about the use of alternative data sources to help address this decline. Third, international organizations such as Eurostat, the United Nations (UN), and the IMF have been particularly active in developing standards and guidelines for data quality, and they have recently focused on the growing use of administrative records and Big Data as requiring revisions to standards that were developed with an exclusive focus on survey and census data.

C. Literature review process

As a first step, Mathematica searched for literature describing data quality standards for integrated data through keyword searches in Google Scholar. Keywords included “Big Data,” alongside “data quality,” “data quality standards,” and “data quality framework.” (“Integrated data” is not yet a commonplace term in the peer-reviewed literature.) We also included articles related to data quality standards for administrative data, such as articles that describe how to extend the Total Survey Error (TSE) framework to administrative data. Date parameters were not necessary because data quality for integrated data, administrative data, and Big Data is a relatively new area of scholarship. Most articles were written after 2000. Specifically for the standards of official statistics, we found the most recently published/uploaded versions.

⁵ Nelson and West (2014) provide a history of the development of population registers in Denmark. Anders and Britt Wallgren, formerly of Statistics Sweden, coauthored the premier text on the use of registers and other administrative data for statistical production (Wallgren and Wallgren 2014).

To cover the grey literature, we also searched major statistical agencies' web sites for data quality standards, including the Australian Bureau of Statistics, Statistics New Zealand, Eurostat, the UNECE, and the UN Global Working Group on Big Data. Additional literature of this nature as well as journal publications was identified after our review had begun by following up on relevant references cited in the materials we read.

We excluded materials that appeared in our search results but on further investigation did not describe data quality standards for integrated data, administrative data, or Big Data. Often, these materials described processes for using administrative data or Big Data with only a mention of "data quality" in the text. We also excluded materials whose principal focus was quality improvement or quality management unless they also addressed quality reporting. Our search resulted in published articles in peer-reviewed journals, such as the *Journal of Official Statistics*; conference proceedings and slides; books and edited volumes; and official documents and whitepapers appearing on statistical organizations' websites. During the review, we further culled the search results to literature that focused on international data quality standards with at least some attention devoted to reporting. We eliminated materials that proved to be off the topic, such as those that merely included a search term but which did not discuss data quality standards or texts that described a case study of combining data sources without discussing quality measurement.

D. Organization of this report

Table I.1 below shows the references that contributed to the central content of the report, broken down by the source country or international organization and type of quality standards covered: general standards, surveys, administrative data, or Big Data. The remainder of the report is organized as follows. Chapter II reviews international standards, beginning with a detailed review of standards issued for the European Union as a whole and then examining

selected countries within Europe and elsewhere and concluding with standards for two international organizations. Chapter III examines a proposal to extend TSE to integrated data, which has been adopted and further developed by the New Zealand statistical authority. Chapter IV reviews key literature on quality assessment focused specifically on uses of administrative data in official statistics. Chapter V does the same for uses of Big Data in official statistics. Chapter VI highlights those findings that most directly address the FCSM working group’s needs with respect to transparency in the reporting of quality for integrated data and presents several conclusions.

Table I.1. Principal sources by country/international organization and type of standards

Country/ International Organization	Quality Standards Type			
	General	Survey	Administrative Data	Big Data
Australia	Australian Bureau of Statistics. (2009) <i>The Australian Bureau of Statistics Data Quality Framework</i> .			Tam, Siu-Ming, and Frederic Clarke. (2015). Big Data, official statistics, and some initiatives by the Australian Bureau of Statistics. <i>International Statistical Review</i> , vol. 83, no. 3, pp. 436-448.
Canada	Brackstone, Gordon. (1999). Managing data quality in a statistical agency. <i>Survey Methodology</i> , 25, 139-149. Statistics Canada. (2017). <i>Statistics Canada’s Quality Assurance Framework</i> . Third edition, 2017.	Statistics Canada. (2009) <i>Statistics Canada Data Quality Guidelines</i> .		

Table I.1. (continued)

Country/ International Organization	Quality Standards Type			
	General	Survey	Administrative Data	Big Data
European Commission/ Eurostat	<p>European Commission. (2011). <i>European Statistics Code of Practice</i>.</p> <p>European Statistical System Committee. (2015). <i>Quality Assurance Framework, Version 1.2</i>. Eurostat.</p> <p>Eurostat. (2015). <i>ESS Handbook for Quality Reports</i>.</p>		<p>Laitila, Thomas, Anders Wallgren, and Britt Wallgren. (2011). <i>Quality Assessment of Administrative Data</i>.</p> <p>Daas, Piet, et al. (2011). <i>Deliverable 4.1: List of Quality Groups and Indicators Identified for Administrative Data Sources</i>, Report for Work Package 4 of the European Commission 7th Framework Program BLUE-ETS.</p> <p>Daas, Piet, and Saskia Ossen. (2011). <i>Deliverable 4.2: Report on Methods Preferred for the Quality Indicators of Administrative Data Sources</i>. Report for Work Package 4 of the European Commission 7th Framework Program BLUE-ETS.</p>	
Finland	<p>Statistics Finland. (2007). <i>Quality Guidelines for Official Statistics</i>. 2nd Revised Edition.</p>			
International Monetary Fund (IMF)	<p>IMF. (2003). <i>Data Quality Assessment Framework and Data Quality Program</i>.</p>			<p>Hammer, Cornelia L., Diane C. Kostroch, Gabriel Quiros, and STA Internal Group. (2017). <i>Big Data: Potential, Challenges, and Statistical Implications</i>. <i>IMF Staff Discussion Note 17/06</i>.</p>

Table I.1. (continued)

Country/ International Organization	Quality Standards Type			
	General	Survey	Administrative Data	Big Data
The Netherlands	<p>Van Nederpelt, Peter. (2009). Checklist Quality of Statistical Output. Statistics Netherlands.</p> <p>Statistics Netherlands. (2014). <i>Quality Guidelines 2014: Statistics Netherlands' Quality Assurance Framework at Process Level.</i></p>		<p>Daas, Piet, et al. (2009). "Checklist for the Quality Evaluation of Administrative Data Sources." Discussion paper 09042. Statistics Netherlands.</p>	
New Zealand			<p>Statistics New Zealand. (2016) <i>Guide to Reporting on Administrative Data Quality.</i></p> <p>Reid, Giles, Felipa Zabala, and Anders Holmberg. (2017). "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." <i>Journal of Official Statistics</i>, vol. 33, no. 2, 477-511.</p>	
OECD	<p>OECD. (2012). <i>Quality Framework and Guidelines for OECD Statistical Activities.</i></p>			
Sweden	<p>Statistics Sweden. (2017). <i>Official Statistics of Sweden—Annual Report 2016.</i></p>			

Table I.1. (continued)

Country/ International Organization	Quality Standards Type			
	General	Survey	Administrative Data	Big Data
United Kingdom	Office for National Statistics. (2013) <i>Guidelines for Measuring Statistical Output Quality</i> . Version 4.1. United Kingdom.		UK Statistics Authority. (2014) "Quality Assurance and Audit Arrangements for Administrative Data."	
	UK Statistics Authority. (2018). <i>Code of Practice for Statistics: Ensuring Official Statistics Serve the Public</i> .		UK Statistics Authority. (2015a). <i>Quality Assurance of Administrative Data: Setting the Standard</i> .	
	Bank of England. (2014). Data quality framework. <i>Bank of England, Statistics and Regulatory Data Division</i> .		UK Statistics Authority. (2015b). <i>Administrative Data Quality Assurance Toolkit</i> .	
United Nations (UN) Big Data Working Group				UN Economic and Social Council. (2015). "Report of the Global Working Group on Big Data for Official Statistics."
UN Economic Commission for Europe			United Nations Economic Commission for Europe. (2011) <i>Using Administrative and Secondary Sources for Official Statistics – A Handbook of Principles and Practices</i> .	United Nations Economic Commission for Europe. (2014). <i>A Suggested Framework for the Quality of Big Data</i> .

II. INTERNATIONAL STANDARDS

Our review of international standards for transparency in the reporting of quality for integrated data begins, appropriately, with the European Union and the work of its two central statistical organizations: Eurostat and the European Statistical System (ESS). Eurostat is charged with the production of official statistics—at the level of all Europe—for the European Union (De Smedt 2016).⁶ The ESS is a partnership between Eurostat and the authorities within each member state of the European Union responsible for statistical production. Under a regulation of the European Parliament, an ESS Committee—consisting of representatives of the member states’ national statistical authorities and chaired by a member of Eurostat—is charged with providing “professional guidance to the ESS for developing, producing, and disseminating European statistics” (<http://ec.europa.eu/eurostat/web/ess/about-us/ess-gov-bodies/essc>). Building on the quality frameworks of European and non-European countries, the ESS developed a quality framework that has become a model for other countries around the globe. While the European framework does not purport to be directed at integrated data, it does acknowledge that some of the estimates produced by European nations may be based on integrated data, and it addresses selected issues raised by such data.

In this chapter we provide a detailed review of the European standards and then follow up with more limited discussions of the standards published by several other countries and international organizations: Canada, Australia, the United Kingdom, the Netherlands, Finland, Sweden, the Organization for Economic Cooperation and Development (OECD), and the IMF.

⁶ Eurostat is a Directorate-General of the European Commission, which is the executive of the European Union. Eurostat is the statistical office of the European Union.

A. European Union

The European Union's approach to the development, assessment, and reporting of quality in official statistics is laid out in three documents: the *European Statistics Code of Practice for the National and Community Statistical Authorities* (European Commission 2011), the *Quality Assurance Framework for the European Statistical System* (ESS Committee 2015), and the *ESS Handbook for Quality Reports* (Eurostat 2015), which includes in an annex the *ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators (QPI)*. We discuss these documents in succession but focus most of our attention on the Handbook, as it addresses most directly the goal of transparency in the reporting of quality.

1. *European Statistics Code of Practice*

The *European Statistics Code of Practice* delineates 15 principles that address: (1) the institutional environment (principles 1 through 6), (2) statistical processes (principles 7 through 10), and (3) statistical output (principles 11 through 15).⁷ The 15 principles are listed in Table II.1.

Assigning these 15 principles to three aspects of the development of statistical estimates is similar to the three-stage approach of the FCSM working group for which this report has been prepared. Where both include statistical processes and statistical output, however, the FCSM working group departs from the ESS in its inclusion of input data quality in lieu of the institutional environment.⁸

⁷ The Code of Practice was first adopted by the ESS Committee in February 2005 and revised in 2011.

⁸ Input data quality, processing quality, and output data quality were the topics of three workshops organized by the FCSM working group between December 2017 and February 2018.

Table II.1. Quality assurance framework of the European Statistical System

Institutional environment	
Principle 1:	Professional independence
Principle 2:	Mandate for data collection
Principle 3:	Adequacy of resources
Principle 4:	Commitment to quality
Principle 5:	Statistical confidentiality
Principle 6:	Impartiality and objectivity
Statistical processes	
Principle 7:	Sound methodology
Principle 8:	Appropriate statistical procedures
Principle 9:	Non-excessive burden on respondents
Principle 10:	Cost effectiveness
Statistical output	
Principle 11:	Relevance
Principle 12:	Accuracy and reliability
Principle 13:	Timeliness and punctuality
Principle 14:	Coherence and comparability
Principle 15:	Accessibility and clarity

Source: European Commission (2011)

For each of the 15 principles, the Code of Practice lists several indicators, which represent ways that national statistical agencies can demonstrate their adherence to or compliance with the principle. These indicators are descriptive of actions that conform to the principle. For example, under the principle of accuracy and reliability there are three such indicators:

- 12.1: Source data, intermediate results and statistical outputs are regularly assessed and validated
- 12.2: Sampling errors and non-sampling errors are measured and systematically documented according to the European standards
- 12.3: Revisions are regularly analyzed in order to improve statistical processes

The brief Code of Practice does not discuss these indicators further. That is left to the *ESS*

Quality Assurance Framework, which we discuss next.

2. *ESS Quality Assurance Framework*

The *ESS Quality Assurance Framework* was produced to assist the national statistical authorities of the member states in implementing the Code of Practice.⁹ Thus the Framework is designed as in aid in *achieving* quality—not in measuring or reporting it. For each of the indicators listed in the Code of Practice, the Framework provides a series of methods at both the institutional level and the product/process level to facilitate achievement of the goal expressed in the indicator.

For example, indicator 12.2 cited above under the principle of accuracy and reliability, states that “sampling errors and non-sampling errors are measured and systematically documented according to the European standards.” At the institutional level, the Framework states that “internal procedures and guidelines to measure and reduce errors are in place and may cover activities such as:

- Identification of the main sources of error for key variables
- Quantification of sampling errors for key variables
- Identification and evaluation of main non-sampling error sources in statistical processes
- Identification and evaluation in quantitative or qualitative terms of the potential bias
- Special attention to outliers as well as their handling in estimation
- Quantification of potential coverage errors
- Quantification of potential measurement errors (comparison with existing information, questionnaire design and testing, information on interviewer training, etc.)
- Quantification of nonresponse errors, including systematic documentation for technical treatment of nonresponse at estimation stage and indicators of representativeness
- Quantification of processing errors
- Analysis of the differences between preliminary and revised estimates”

⁹ Adherence to the Code of Practice is monitored through periodic peer reviews of the national statistical authorities. The first round was conducted between 2006 and 2008. A second round was initiated in December 2013. See <http://ec.europa.eu/eurostat/web/quality/peer-reviews>.

At the product/process level, three methods are listed:

- Periodic quality reporting on accuracy is in place (serving both producer and user perspectives)
- Quality reporting on accuracy is guided by ESS recommendations (for example, *ESS Handbook for Quality Reports*)
- Methods and tools for preventing and reducing sampling and non-sampling errors are in place

The methods at the institutional level provide more guidance in the reporting of quality than do those at the product/process level. However, the second of the three methods at the product/process level explicitly addresses quality reporting and refers to the *ESS Handbook for Quality Reports*. This document is more on target than the *Quality Assurance Framework* with respect to standards for transparency in the reporting of quality. For that reason we turn now to an extended discussion of the Handbook.

3. *ESS Handbook for Quality Reports*

The express purpose of the Handbook is to provide guidance to national statistical authorities in “the preparation of comprehensive quality reports for a full range of statistical processes and their outputs” (Eurostat 2015).^{10, 11} Statistical processes include, for example, sample surveys, censuses, and statistical uses of administrative data.

Specific objectives of the guidelines presented in the Handbook are:

- To promote harmonized quality reporting across statistical processes and their outputs within a Member State and hence to facilitate comparisons across processes and outputs;
- To promote harmonized quality reporting for similar statistical processes and outputs across Member States and hence to facilitate comparisons across countries; and

¹⁰ For an example of a recent quality report prepared for the European Union see Eurostat (2017).

¹¹ The quality reports discussed in the Handbook have their closest analog in the U.S. in the quality profiles that have been prepared for a variety of federal datasets; see Kasprzyk and Kalton (2001). We return to the subject of quality profiles in Chapter VI.

- To ensure that reports include all the information required to facilitate identification of statistical process and output quality problems and potential improvements.

The guidelines address each of eight Code of Practice principles, including the five dimensions of statistical output quality (principles 11 through 15) and three additional principles, involving confidentiality (principle 5), burden (principle 9), and cost (principle 10). The Handbook also includes guidelines for assessing and reporting on statistical processing, which does not correspond to any of the principles. The Handbook notes that because quality at the institutional or statistical process stage bears directly on output quality, quality assessments for all 15 principles would involve some redundancy. Limiting the quality assessment to the five dimensions of statistical output quality plus three principles that are not so clearly reflected in output quality—and adding statistical processing—is considered sufficient for a comprehensive assessment of quality. Below we summarize the guidelines for quality reporting for each of these eight principles plus statistical processing.

Included in the recommendations for quality reporting for the five dimensions of statistical output quality are 16 quantitative indicators. These include common measures of survey data quality as well as measures of other aspects of data quality not typically quantified in the U.S. The full set of indicators is listed below, where the prefix R stands for relevance, A for accuracy, TP for timeliness and punctuality, CC for coherence and comparability, and AC for accessibility and clarity:

- R1. Data completeness rate, which can be calculated for a given dataset and time period and is defined as the ratio of the number of data cells reported to the number of data cells required (by Eurostat or the relevant statistical agency)
- A1. Sampling error indicators: the coefficient of variation and the confidence interval of an estimate
- A2. Overcoverage rate, defined as the proportion of units accessible via the frame that do not belong to the target population (are out of scope)

- A3. Common units proportion, defined as the proportion of units in the survey covered by an administrative source
- A4. Unit nonresponse rate, defined as the proportion of eligible (in-scope) units with no information or no usable information and calculated either weighted or unweighted
- A5. Item nonresponse rate, defined for a given item as the ratio of the number of in-scope units that have not responded relative to the number required to respond to that item
- A6. Data revision average size, defined as the average difference between a later and an earlier estimate of a key item
- A7. Imputation rate, defined as the ratio of the number of imputed values to the total number of values requested for that variable and calculated either weighted or unweighted
- TP1. Time lag for first results, defined as the length of time between the end of the event or phenomenon they describe and their availability
- TP2. Time lag for final results, defined as the length of time between the end of the event or phenomenon they describe and their availability
- TP3. Punctuality, defined as the time lag between the delivery or release data of data and the target date announced in an official release calendar, specified in regulations, or agree among partners
- CC1. Asymmetry for mirror flows statistics, defined as the difference between inbound and outbound flows (for example, between countries) divided by the average of the two flows
- CC2. Length of comparable time series, defined as the number of reference periods in a time series since the last break in the series
- AC1. Data tables consultations, defined as the number of times users consulted a particular data table, where multiple views within a single session count as one view
- AC2. Metadata consultations, defined as the number of times users viewed metadata within a statistical domain
- AC3. Metadata completeness rate, defined as a ratio of the number of metadata elements provided to the total number of applicable elements

A companion document on ESS quality and performance indicators, included as an annex (or appendix), provides detailed instructions for computing each of these indicators as well as guidance in their use and interpretation. In our discussion of quality reporting below we highlight the indicators that apply to each of the five dimensions of statistical output quality.

The Handbook also identifies six types of statistical processes that may have been used to generate the statistical output whose quality is the subject of the report:

- Sample surveys

- Censuses
- Statistical processes using administrative sources
- Statistical processes involving multiple data sources
- Statistical processes for generating price and other economic indexes
- Statistical compilations (such as economic aggregates)

The discussion of quality for the accuracy and reliability dimension differentiates among all six types of statistical processes. The discussion of quality for the relevance dimension differentiates among three of the six. There is no differentiation among these statistical processes for the three remaining dimensions of output quality and the three principles of confidentiality, burden, and cost. However, there is a general recommendation that whenever multiple data sources were used—particularly different types of data sources, such as a sample survey and administrative records—a separate quality report should be produced for each data source and not just the combination of multiple data sources. This point is made in the next chapter as well.

a. Relevance

The dimension of relevance is focused on the users of the statistical outputs and to what extent the data can be shown to satisfy their needs. To assess relevance the quality report should include:

- A classification of users
- A breakdown of the uses for which different groups of users need the outputs and the key outputs that address each group's needs
- The statistical authority's priorities in addressing these needs
- Discrepancies between the operational concepts used in generating the data and the ideal concepts from the perspective of users
- The degree of completeness of the data with respect to required contents as defined by the ESS or other international guidance
- An account of how the information on user needs was obtained

Examples of what might be included in this last item are advisory committees, user groups, ad hoc focus groups, user surveys, and complaints.

Some additional requirements are imposed when the statistical process includes administrative data or the outputs are price indices or statistical compilations. For the former the quality report should explicitly compare the concepts or definitions embedded in the data—which are fixed—to those desired by key users. For price indices the quality report should discuss issues related to defining and operationalizing the target of estimation, as an exact specification is rarely possible. Citing practices recommended in international documents to support the choice of methods provides a quality standard under these circumstances. Likewise, relating the methods used in preparing statistical compilations such as National Accounts to international guidelines or other consensus is a key element in assessing the quality of such statistics.

The one quality and performance indicator for relevance, R1, is the data completeness rate, or the ratio of the number of data cells provided to the number required.

b. Accuracy and reliability

Quality assessment of the accuracy and reliability dimension involves so many facets that nearly half of the Handbook's main text is devoted to that topic. Accuracy is broken down into overall accuracy, sampling error, and non-sampling error. Non-sampling error is further divided into four sub-concepts: (1) coverage error, (2) measurement error, (3) nonresponse error, and (4) processing error. These forms of non-sampling error apply not just to probability surveys but to other types of statistical processes as well, but the authors of the Handbook note that the meanings of these error sources are not as well established for these other domains.

As noted above, the guidelines for quality reporting on accuracy and reliability distinguish among the six types of statistical processes listed earlier. There is a separate, preliminary

discussion that applies to all statistical processes. Additional discussion covers some general issues in the reporting of accuracy that occur across multiple types of statistical processes. These include model assumptions and associated errors, seasonal adjustment, imputation, mistakes, and revisions. The reporting of accuracy for statistical processes using administrative sources and multiple data sources speaks most directly to the needs of the FCSM working group. Therefore, we review the guidelines for quality reporting for these statistical processes in greater detail than we do the other processes.

All statistical processes. The Handbook's discussion of quality reporting on accuracy that should be applied to all statistical process distinguishes between random error, which tends to cancel out on average, and systematic error, which introduces bias. The Handbook encourages an assessment of the risks of bias, which can be expressed in quantitative or qualitative terms. Qualitative assessments of bias should include not only the likely sign of bias but an estimate of its general magnitude and the basis for this assessment. As a general reference on the reporting of accuracy, the Handbook cites the FCSM's Statistical Policy Working Paper 31, *Measuring and Reporting Sources of Error in Surveys* (FCSM 2001).

Quality reports for all statistical processes should include:

- Identification of the chief sources of error for the main variables
- If micro data are made accessible for research purposes, additional information that may assist users, if such information is considered essential
- A summary assessment of all sources of error with special focus on the key estimates
- An assessment of the potential for bias (sign and order of magnitude) for each key indicator in quantitative or qualitative terms—as discussed in the preceding paragraph

These requirements apply to assessments of overall accuracy.

Sample surveys. More than half of the discussion of accuracy and reliability is devoted to sample surveys, which reflects the attention that has been focused on error in this method of data

collection.¹² Recommendations for quality reporting for sampling error and non-sampling error are presented separately.

The discussion of quality reporting of sampling error includes not just probability samples but non-probability samples as well, although the types of non-probability samples considered are very limited. The single quality and performance indicator for sampling error is the standard error, but the Handbook makes several additional recommendations for reporting on sampling error. It is suggested that sampling error should be presented not just for estimates of level but estimates of change as well. There is flexibility in the form that the presentation of sampling error should take (for example, coefficients of variation versus confidence intervals). The treatment of outliers should be described. For non-probability samples, cut-off sampling is distinguished from other forms of sampling, with different prescriptions for reporting.

The discussion of non-sampling error includes coverage errors, defined as divergences between the frame and target populations; measurement errors; nonresponse errors; and processing errors. Coverage errors include undercoverage, overcoverage (inclusion of units not in the target population, such as deceased persons), and duplication. One indicator, A2, the overcoverage rate, should be included in the quality report. Undercoverage, the Handbook notes, is the most challenging to measure. Several ways of assessing coverage error of this type are discussed, including comparison with external data, re-interviews, experiments, and comparison of reported and edited values. No specific indicators are offered for undercoverage, however.

There is no performance indicator for measurement error. To address this form of non-sampling error the Handbook recommends the following:

- Identification and general assessment of the main risks of measurement error

¹² Witness in particular the TSE model as presented in Groves et al. (2009) and the subsequent research that it has generated.

- Assessments based on comparisons with external data, re-interviews, experiments, or data editing, depending on what is available
- The efforts made in questionnaire design and testing, information on interviewer training, and other work on error reduction
- Attachment of questions in an appendix (or by hyperlink if their length is excessive)

On this last point, there is no discussion of the challenges presented by computerized instruments, where a questionnaire, per se, does not exist.

The discussion of nonresponse error focuses on the calculation and reporting of unit and item nonresponse rates—the performance indicators A4 and A5. In addition the quality report should include:

- A breakdown of non-respondents according to the cause for nonresponse
- A qualitative statement on the bias risks associated with nonresponse
- Measures taken to reduce nonresponse
- Technical treatment of nonresponse at the estimation stage

Processing errors arise from coding and editing (although the more important impact of editing may be in reducing measurement error), and while no indicators are provided, it is noted that the estimation of coding errors requires some type of repeated coding. The quality report should identify the main issues regarding processing errors for the statistical process and its outputs and, where relevant and available, provide an analysis of processing errors affecting individual observations. Absent the latter, a qualitative assessment should be presented.

Censuses. For censuses the error sources are similar to those for sample surveys, except that sampling error is generally not a consideration.¹³ In its place, coverage error becomes the primary focus. Measurement and nonresponse error can be important as well, but we note that, in

¹³ When a census includes a sample survey of households responding to the census, as the U.S. Census did for a number of years until 2010, the sample component can be treated as a sample survey but with a frame shared with the census—as well as coverage error.

the U.S. at least, the variables collected in the census tend to have lower error from these sources than variables that are reserved for household sample surveys—such as income.

The quality and performance indicators A2, A4, and A5 are applicable to censuses as well.

In addition, quality reports should include:

- An assessment of undercoverage and overcoverage
- A description of methods used to correct for undercoverage and overcoverage
- A description of methods and an assessment of the accuracy if a cut-off threshold is used in place of collecting data from all units
- An evaluation of measurement error
- An evaluation of nonresponse error
- An evaluation of processing error

Processing error includes data entry errors and, if applicable, coding errors.

Statistical processes using administrative sources. The discussion of administrative sources focuses on register-based statistics and begins by defining three types: (1) estimates produced from one register, (2) integration of several registers in order to obtain and describe new populations and variables, and (3) event-reporting systems. Examples of this last type include systems that capture reports of crimes and vehicular accidents.

Registers, it is noted, cover the universe of units meeting a particular definition, so sampling errors do not exist. However, potential errors for estimates based on a single register derive from over- or undercoverage, which may be attributable to lags in entering information into registers; nonresponse, which includes missing data at both the unit and item levels; measurement error; processing errors, which may be due to the provider and/or the statistical agency, if separate; and conceptual differences between the register and target, including those derived from multi-valued variables (for example, businesses with activity in more than one industry or persons with more than one job).

Integration of multiple registers necessarily involves record linkage of some variety. The quality of the linkage depends in large part on the quality of the unit identifiers in each register. Consequently, assessing the accuracy of these identifiers before using them for linkage is important. Linkage errors in the form of false matches or false nonmatches are the greatest risk when integrating registers.

Data quality in event-reporting systems depends primarily on the completeness of reporting. Classification error in recording the type of event is described in the Handbook as a processing error although it would seem that the basic reporting of events can readily introduce such errors.

Statistical processes involving multiple data sources. When statistical processes involve multiple data sources, the individual sources (for example, surveys, censuses, administrative records) should be assessed as described above, but an assessment of the “whole picture” as well as the individual components is necessary. A quality report for a statistical process involving multiple data sources should include an overall description of the how the process is organized, the various segments that are included, and a summary of the quality aspects.

For assessing the quality of the final product the sole suggestion that the Handbook provides applies only to estimates that are produced in preliminary form with subsequent revisions. For such estimates the Handbook recommends assessing the typical amount of revision. Small revisions may be indicative of high accuracy under the assumption that the successive estimates are converging on the true value. Of course, this does not address the error in the initial estimates. Such error may not have been reduced if the successive estimates show little change. Even if the revisions do exhibit marked change, one or more of the individual error components in the initial estimates may not have been reduced.

Performance indicators A1 through A5 are listed as applicable (and A6 if revisions are part of the statistical process) although only one of these explicitly includes multiple data sources: the

common units proportion, which is defined as the proportion of survey units for which there is a corresponding administrative unit. Presumably this calculation could be restricted to applicable survey units, although defining applicable units operationally may be difficult. To use an example from the U.S., the number of Supplemental Nutrition Assistance Program (SNAP) administrative units could be expressed as a proportion of the total number of units in the survey sample, but that would seem to be of less interest than expressing the number of SNAP administrative units relative to the number of sample units that represent the target population for SNAP.

Price and other economic index processes. Price indices are commonly constructed from multiple sources of data. A quality report should assess each component. As these indices necessarily involve sampling from multiple universes (households or companies; products), and not all of the sampling may be probability-based, the assessment of sampling error is important. Yet there is no generally agreed-upon approach. Nevertheless, all relevant sampling dimensions should be discussed in a quality report. Coverage error and, more generally, limitations in coverage in each of the dimensions of sampling should be discussed as well. It is noted that when non-probability sampling is used, the distinction between coverage error and sampling error tends to become blurred.

Quality adjustment—changes over time in the product mix—is a particularly important source of error in price indices. Quality reports should address this source of error as a measurement problem. Nonresponse and other sources of errors are often considered secondary in importance to sampling, coverage, and quality adjustment, but should be assessed nonetheless.

Statistical compilations. Statistical compilations include most prominently economic and other aggregates. Different approaches are taken to the assessment of accuracy for such compilations because such assessment is challenging. Given that the use of such statistics often

focuses on change over time, consistency in the error contributions from sources that cannot be directly measured tends to reduce their importance. As with other economic indicators where direct measurement of error may not be possible, the analysis of revisions becomes an important alternative. Measurement of the non-observed economy is an especially challenging problem in that such measurement cannot rely on the usual administrative and survey sources. For Europe the OECD has produced a handbook on measuring this component of each national economy.

General issues. The Handbook provides additional discussion of issues in the reporting of accuracy that occur across multiple types of statistical processes. These include model assumptions and associated errors, seasonal adjustment, imputation, mistakes, and revisions.

Modeling may play a role in many statistical estimates. Model-assisted sampling is sometimes used to improve the precision of sample estimates, but its impact is reflected in the results of variance calculations, so no separate assessment is needed. Model-dependent estimation, however, requires separate discussion. When modeling is used to address particular sources of error, the modeling assumptions should be discussed in quality reports along with those error sources. When the target estimation is model-based, the model should be detailed in the quality report, and its validity for the data to which it is applied should be assessed.

Seasonal adjustment is, of course, heavily model-dependent. The ESS has developed a set of guidelines on seasonal adjustment, which Eurostat has adopted. The guidelines include a metadata template, which, if completed, can be referenced in the seasonal adjustment section of the quality report. The following should be provided in addition:

- A short description of the method used, including pre-treatment (calendar effects corrected for, calendar used, type of outliers detected and corrected, model selection and revision, and decomposition scheme adopted) and specification of the seasonal adjustment tool chosen (software, its version and operating system)
- Specification of the quality measures and diagnostics used to validate the identified model and the results of the seasonal adjustment process

- The approach chosen for handling revision of seasonally adjusted data in combination or not with revision of raw data

In the absence of a completed metadata template, a fuller description covering each item of the seasonal adjustment guidelines should be provided.

Imputation should also be covered in the quality report. We note that the imputation rate is one of the quality indicators listed above. The quality report should include imputation rates as a way of documenting the extent to which imputation was used in producing the final data product. The report should also contain a discussion of the method(s) of imputation and what is known about the effects of imputation on the estimates—including variability. Imputation can be discussed under the source of error that it is intended to reduce, rather than separately.

Error can also be introduced by mistakes in processing. The Handbook advises that when processing errors or mistakes in calculation or presentation (such as publishing the wrong numbers in a table or press release) create the need for subsequent revisions, the errors should be documented when revisions are released. More generally, quality reports should include discussion of steps taken to minimize the risk of serious mistakes and how they are handled if discovered.

The use of the magnitudes of revisions as a way of assessing error in the production process has been discussed, but revisions are also a distinct topic for inclusion in a quality report. The Handbook notes that the Code of Practice requires that revisions follow “standard, well-established and transparent procedures.” This includes both planned and unplanned revisions. The quality report should describe the revision policy, present the number of revisions, give the average size of revisions for one or more measures, outline the main reasons for revisions, and document the extent to which revisions have improved accuracy.

c. Timeliness and punctuality

Timeliness refers to the length of time between the reference period for a statistical estimate or dataset and when it is made available to users; punctuality refers to whether data were delivered on the data they were scheduled for release. Both are straightforward to measure and are captured in three indicators reported earlier:

- TP1. Time lag for first results, defined as the length of time between the end of the event or phenomenon they describe and their availability
- TP2. Time lag for final results, defined as the length of time between the end of the event or phenomenon they describe and their availability
- TP3. Punctuality, defined as the time lag between the delivery or release data of data and the target date announced in an official release calendar, specified in regulations, or agree among partners

The Handbook adds that the quality report should explain the reasons for non-punctual data releases.

d. Coherence and comparability

The Handbook defines six types of coherence and comparability:

- Coherence across domains, or the extent to which statistics can be reconciled with those that were obtained for other statistical domains or through other data sources
- Coherence between sub-annual and annual statistics
- Coherence with National Accounts
- Coherence internally, or the extent to which statistics are consistent within the same dataset
- Comparability geographically, or the extent to which there is comparability across geographic areas, which for Eurostat includes comparability across countries
- Comparability over time, or the extent to which statistics are either comparable or can be reconciled over time

The Handbook underscores the importance of being able to combine and make joint use of related data derived from different sources.

A lack of coherence or comparability may derive from differences in concepts or methods.

Differences in concepts may apply to the target population, geographic coverage, reference

period, or definitions of data items and classifications. Differences in methods may involve the choice of frame population; sources of data; sample design; procedures for data collection, capture or editing; or imputation and estimation. The Handbook inserts a cautionary note that accuracy and coherence/comparability are easily confounded. That is, a seeming inconsistency or lack of comparability may be due to sampling error or other source of inaccuracy.

Only two quality and performance indicators are defined for coherence and comparability:

- CC1. Asymmetry for mirror flows statistics, defined as the difference between inbound and outbound flows (for example, between countries) divided by the average of the two flows
- CC2. Length of comparable time series, defined as the number of reference periods in a time series since the last break in the series

The small number of indicators belies how much information on coherence and comparability should be included in a quality report. The following are requested:

- Brief descriptions of conceptual and methodological metadata elements that could affect coherence or comparability
- An assessment of the possible effect of each reported difference on the output values
- Differences between the statistical processes employed and the corresponding European or international regulation or standard
- A quantitative assessment of comparability across regions
- A coherence/comparability matrix defined at the ESS level summarizing by region the possible sources of a lack of comparability relative to a specified standard
- An assessment of any discrepancies in mirror statistics
- Location of breaks in series and their reasons and how they are being treated
- Comparisons with National Accounts where relevant and feedback from the producers of National Accounts with regard to coherence and accuracy issues
- Any lack of internal coherence in the output of the statistical process

The Handbook provides extensive examples bearing on the types of assessments that may be made, underscoring the importance assigned to coherence and comparability.

e. Accessibility and clarity, dissemination format

Accessibility and clarity refer to the simplicity and ease of use of the data, including how and under what conditions users can access the data and how readily users can correctly interpret statistics in light of the supporting information and other assistance that is provided. The dissemination format refers to how the statistical data and metadata are distributed to users, including the medium and format. Types of dissemination include news releases, publications, on-line databases, micro data access, and other forms—and, if applicable, their pricing. Also relevant to dissemination are the ways in which documentation on methodology and quality are made available.

The quality report should differentiate among types of users and how well their differing needs have been addressed. User feedback is the best source of information for responding to this aspect of the report.

There are three quality and performance indicators:

- AC1. Data tables consultations, defined as the number of times users consulted a particular data table, where multiple views within a single session count as one view
- AC2. Metadata consultations, defined as the number of times users viewed metadata within a statistical domain
- AC3. Metadata completeness rate, defined as a ratio of the number of metadata elements provided to the total number of applicable elements

In addition to presenting these the report should provide a description of the conditions of access to the data; a summary description of the metadata that accompanies the data, distinguishing between what is provided for less sophisticated users versus more advanced users; and a summary of feedback received from users on each of accessibility, clarity, and the dissemination format.

f. Cost and burden

The Handbook recommends that a quality report should include the following with respect to the cost principle:

- Annual operational cost with a breakdown by major cost components
- Recent efforts to improve cost-efficiency
- The procedures for internal assessment and for independent external assessment of efficiency
- The extent to which routine operations—for example, data capture, coding, validation, and imputation—are automated
- The extent to which information and communications technology is used effectively for data collection and dissemination and a discussion of the improvements that could be made

The challenges associated with obtaining a breakdown of costs by their major components are acknowledged; nevertheless, having such a breakdown is critical to the development of strategies to reduce costs and improve efficiency.

With respect to respondent burden the Handbook recommends that a quality report include the following:

- Annual respondent burden in financial terms and/or hours
- Reduction targets for respondent burden
- Recent efforts to reduce respondent burden
- Whether the range and detail of data collected by survey is limited to what is absolutely necessary
- The extent to which data sought from businesses is readily available from their accounts
- Whether electronic means are used to facilitate data collection
- Whether best estimates and approximations are accepted when exact details are not readily available
- Whether reporting burden on individual respondents is limited to the extent possible by minimizing the overlap with other surveys

On the first point the Handbook notes that, in principle, the financial cost of the burden imposed on respondents in completing a questionnaire can be calculated as the product of the number of

respondents, the average time required to assemble and enter the information or participate in an interview, and the average hourly cost of the respondent's time. Because of the difficulty of determining this last component, burden is often calculated as the product of just the first two components.

g. Confidentiality

In discussing confidentiality, the Handbook distinguishes between policy—the legislative or other measures prohibiting unauthorized disclosure of data that identify a person or economic entity—and data treatment, the procedures that are applied to data to ensure “statistical confidentiality” and prevent unauthorized disclosure. Building on this the Handbook recommends that a quality report include:

- Whether or not confidentiality is required by law and, if so, whether survey staff have signed legal pledges to maintain the confidentiality of the information they collect or process
- Whether external users may access micro data for research purposes and, if so, the confidentiality provisions that are applied
- The procedures for ensuring confidentiality during collection, processing, and dissemination, including rules for determining confidential cells in output tables and procedures for detecting and preventing residual disclosure

Not mentioned here or in the longer discussion are the measures taken to confirm the effectiveness of the statistical procedures employed to prevent unintentional disclosure in the tables and micro data released to the public, but that is something with which a national statistical office would have to address in asserting that the data released are “safe.”

h. Statistical processing

Statistical processing in the Handbook encompasses all of the operations that are performed on data to derive new information in accordance with a given set of rules. Statistical processing encompasses the following elements:

- Source data
- Frequency of data collection

- Data collection
- Data validation
- Data compilation
- Adjustment

Guidelines for describing each of these components are presented, but the only quality and performance indicator requested is the imputation rate, A7, which is one of the indicators of accuracy discussed earlier.

B. Selected national statistical organizations

We would not expect other members of the European Union to add much if anything to what is already covered by Eurostat and the ESS, but we include discussions of the United Kingdom and the Netherlands to highlight unique features of their quality frameworks. We also include Finland and Sweden to show how the Eurostat and ESS standards are incorporated into their own quality frameworks. We begin, though, with a discussion of quality standards for Canada and Australia. New Zealand is discussed at length in the next chapter because of its explicit acknowledgment of the demands presented by integrated data. Finally, we close with a discussion of the quality frameworks of the OECD and the IMF. As seen below, most other countries have built off of or in tandem with the standards developed by Eurostat and the ESS.

1. Canada

Statistics Canada's Quality Assurance Framework (Statistics Canada 2017)¹⁴ reflects the agency's mission statement, "Serving Canada with high-quality statistical information that matters." Like many other national statistical agencies and international organizations, Statistics

¹⁴ The Quality Assurance Framework was first released in 1997 and updated in 2002. The 2017 release is the third edition, which "was inspired by the generic National Quality Assurance Framework template developed by a United Nations Statistics Division Expert Group. In particular, this version expands the scope of the Statistics Canada QAF by discussing quality management in the Agency's corporate environment and statistical programs."

Canada defines the quality of its official statistics in terms of their fitness for use. Underlying its strategies for effective quality management of statistical information are eight guiding principles:

- Quality is multi-dimensional
- Quality is relative, not absolute
- Every employee has a role to play in assuring quality
- Quality must be built in at each phase of the process
- Balancing the dimensions of quality is best achieved through a team approach
- Quality assurance measures must be adapted to the specific program
- Users must be informed of data quality so that they can judge whether the statistical information is appropriate for their particular use
- Quality assurance is a continuous practice

The emphasis on transparency embodied in the seventh principle underscores a user focus in the agency's approach to the production of official statistics.

Reflecting Statistics Canada's multi-dimensional view of quality, the agency defines the quality of its statistical information and assesses its fitness for use with respect to six dimensions:

- Relevance, which reflects the degree to which statistical information meets user needs
- Accuracy, which reflects the degree to which statistical information correctly describes the phenomena it was designed to measure
- Timeliness, which refers to the delay between the end of the reference period to which statistical information pertains and the date on which the information becomes available
- Accessibility, which refers to the ease with which statistical information can be obtained
- Coherence, which reflects the degree to which statistical information is logically consistent and can be brought together with information from other sources or different time periods
- Interpretability, which reflects the availability of supplementary information (metadata) necessary to understand, analyze, and utilize the statistical information appropriately

In discussing the principle on informing users of data quality, the agency notes that some of these dimensions can be observed directly by the user (timeliness, for example), but for most of the others, the user requires objective information for which the agency may be the sole source.

The six dimensions of quality are discussed in depth in a section of the Quality Assurance Framework on statistical outputs. For each of these dimensions, the document provides a detailed description, summarizes how it is assessed, and lists a number of initiatives that Statistics Canada has undertaken to promote the dimension in its statistical programs. For example, to promote accuracy, Statistics Canada:

- Incorporates quality assurance measures into program and process design, implementation and execution
- Manages and monitors accuracy during implementation and execution of its statistical programs and processes
- Assesses accuracy and reliability, both pre-release and post-release, and communicates the results

Implementation of this last initiative includes periodic compilation and dissemination of quality reports, which include both quantitative and qualitative analyses of all types of errors.

Statistics Canada has been a world-wide leader in the use of administrative records as an alternative to the collection of data from survey respondents. The respondents to Statistics Canada's major household surveys may allow the agency's use of their administrative data on income and participation in government programs in place of responding to questions on these topics. This type of data integration is one that has been highlighted by advocates of greater use of integrated data in the U.S. The *Statistics Canada Quality Guidelines* (Statistics Canada 2009), which predate the Quality Assurance Framework,¹⁵ focus on censuses and sample surveys, but the Guidelines extend the term "survey" to encompass "any activity that collects or acquires statistical data," which includes not only censuses and sample surveys but collections of data from administrative records and statistical activities, "in which data are estimated, modeled, or

¹⁵ The latest version of the Guidelines is the fifth edition. The second edition was published in April 1987. The date of the first edition was not reported.

otherwise derived from existing statistical data sources.” Thus, the Guidelines include a chapter on the use of administrative data, which we discuss next.

Administrative records exist to serve an administrative purpose, not a statistical one. Potential statistical uses were not considered in most cases when the administrative program was established, and the statistical agency may have little ability to influence the content of another agency’s administrative data. Consequently, the Guidelines advise that any decision to use administrative records in conjunction with a survey “must be preceded by an assessment of such records in terms of their coverage, content, concepts and definitions, the quality assurance and control procedures put in place by the administrative program to ensure their quality, the frequency of the data, the timeliness in receiving the data by the statistical agency and the stability of the program over time.” In assessing the quality of administrative data, the quality dimensions of relevance, accuracy, timeliness, and coherence all merit serious consideration.

The Guidelines note issues that may arise in combining administrative data with survey data.

The Guidelines include cautions along with recommendations:

- If administrative data are used as a frame in addition to or in place of another one obtained from data collection, it may not be possible to analyze the issues of coverage and nonresponse
- Indicate the contribution to key estimates from administrative data
- If administrative data are used as a frame, and some elements have been imputed, report the imputation rate for unit or item nonresponse and explain how the imputation was performed
- If the administrative data are summed to produce a statistical output, include an estimate of the loss of precision due to imputation
- If administrative data make up part of an estimate, with the rest derived from survey data, report the portion of the frame covered by administrative data as well as the portion of the estimate
- Produce a response rate combining both the administrative portion and the survey portion as explained in Trepanier et al. (2005)

Other issues that have arisen in the context of Statistics Canada's experience in combining administrative data and survey data are discussed by Lavallee (2000, 2005).

Statistics Canada was also a pioneer in record linkage. The theory underlying probabilistic record linkage was given its mathematical foundation in a 1969 paper by Ivan Fellegi and Alan Sunter (Fellegi and Sunter 1969) of Statistics Canada, then called the Dominion Bureau of Statistics. Statistics Canada also produced one of the first software packages to apply these methods. The Guidelines discuss issues and provide a number of recommendations for situations when the use of administrative records requires record linkage, whether by exact matching (as is likely to be the case when administrative records are used in place of survey responses) or probabilistic methods (although that term is not used).

Statistics Canada has a formal Policy on Informing Users of Data Quality and Methodology, which was approved March 31, 2000.¹⁶ The Policy evokes the transparency theme that the FCSM working group is seeking to articulate. In addition to providing standards and guidelines, the policy document lays out several general principles that should govern their implementation:

- Users must be provided with the information necessary to understand both the strengths and limitations of the data being disseminated.
- The documentation provided to users on data quality should engender an awareness of quality as an issue in the proper use of the data.
- The documentation on methodology must permit users to assess whether the data adequately approximate what they wish to measure, and whether the estimates were produced with tolerances acceptable for their intended purpose.
- The documentation provided should be clear, well organized and accessible. Accuracy indicators should not be technically difficult for the intended clientele to understand or use.
- The descriptions of methodology and the indicators of data accuracy should be carefully integrated whenever this will benefit the user's understanding.
- Specific standards for the level of detail to be provided in documentation on data quality or methodology (listed in the document) are mandatory but minimum requirements.

¹⁶ A revision was issued November 25, 2002 (Statistics Canada 2002).

- The detail and frequency of the updating of the documentation on data quality for purposes of the Policy should consider
 - The intended uses of the data;
 - The potential for error and its significance to the use of the data;
 - Variation in accuracy and coherence over time;
 - Cost of the evaluation of data quality relative to the overall cost of the statistical program;
 - Potential for subsequent improvement of quality and efficiency;
 - Applicability and utility of the indicators of accuracy to users.

The standards included in the Policy specify the inclusion of a number of descriptive statements about the data sources and methodology, the concepts and variables measured, and data accuracy. Quantitative measures are limited to estimates of sampling error, response rates, and imputation rates. The guidelines cover additional documentation that may be of benefit to users but do not extend to reporting on the quality of statistical estimates derived explicitly from integrated data.

2. Australia

The *Australian Bureau of Statistics (ABS) Data Quality Framework*, issued in May 2009, is presented as ABS's official data quality framework for all statistical products, applying to administrative data as well as survey-based products (ABS 2009). The ABS Data Quality Framework is based on *Statistics Canada's Quality Assurance Framework* and the *European Statistics Code of Practice*. As such, it shares many elements from these two sources. Similar to other national statistical agencies, the ABS defines quality as "fitness for purpose," which implies both an assessment of an output and a reference to its intended application.

The ABS Data Quality Framework comprises seven dimensions: (1) the institutional environment, which refers to context factors that might impact credibility and which is where the ABS includes privacy/confidentiality aspects of data, (2) relevance, (3) timeliness, (4) accuracy,

(5) coherence, (6) interpretability, and (7) accessibility. Each of these dimensions is further defined by key aspects and a number of suggested questions to assess the dimension.

As an example, the relevance dimension, which represents how well a statistical product meets the needs of users, can be evaluated by the following aspects:

- Scope and coverage
- Reference period
- Geographic detail
- Main outputs or data items
- Classifications and statistical standards (their conformance with target concepts)
- Type of estimates available
- Other cautions

Suggested questions to assess the relevance of a statistical output include:

- About whom, or what, were the data collected?
- How useful are these data at small levels of geography?
- Does this data source provide all the relevant items or variables of interest? Does the population represented by the data match the data need?
- To what extent does the method of data collection seem appropriate for the information being gathered?
- If rates and percentages have been calculated, are the numerators and denominators consistent?

We note in particular the inclusion of geographic detail, one form of granularity, which the CNSTAT panel elevated to a proposed dimension.

The importance of each dimension will vary depending on the data source and research context, although the application of this principle is left somewhat vague, with the ABS noting, “We recommend that judgment is used in making assessments of quality, and that the quality dimensions are evaluated appropriately for the particular context.” For example, if a key purpose of a statistical product is to facilitate comparing and contrasting estimates, the dimension of

coherence will assume elevated importance. Likewise, traditional survey-based measures of statistical accuracy may not apply to administrative data, in which case timeliness or relevance may assume greater importance. The Framework seems more geared toward descriptive versus evaluative reporting of quality, with judgments about “good versus bad” or “high versus low quality” being left to the user to assess, based upon their specific needs.

The ABS recommends the development of a “quality statement” as an aid in assessing the quality of a statistical product. A quality statement should follow the guidance of the Framework in communicating key characteristics of the data that may affect their quality and should include both strengths and limitations. Quality statements can vary in their level of detail. The ABS has produced succinct summaries called “quality declarations,” which present key information about the quality of the data in statistical releases and may include links to more detailed information. The ABS notes, however, that quality declarations are not intended to substitute for more comprehensive quality statements.

3. United Kingdom

Multiple documents from the UK speak to data quality in official statistics. Three are discussed here. Three additional documents, dealing exclusively with administrative records, are discussed in Chapter IV.

The major text on data quality in the UK is the *Code of Practice for Statistics* (UK Statistics Authority 2018), which is a set of quality principles and guidelines developed by the UK Statistics Authority and which adheres to the United Nation’s *Fundamental Principles of Official Statistics* and the *European Statistics Code of Practice*. The UK code’s purpose is “to ensure: that the range of official statistics meets the needs of users; that the statistics are produced, managed and disseminated to high standards; and that the statistics are well explained.” To this

end, the code establishes eight principles with associated practices, or processes, to achieve each.

The eight principles are:

1. Meeting user needs refers to statistics meeting the requirements for informed decision making by government, public services, business, researchers, and the public.
2. Impartiality and objectivity means that official statistics should include information about processes, which should be managed objectively.
3. Integrity means that all stages in the production, analysis, and dissemination of official statistics should be free from political or personal interests.
4. Sound methods and assured quality refers to maintaining scientific best practices and monitoring quality throughout the process of producing and releasing official statistics.
5. Confidentiality means private information about individual persons should be confidential and used for statistical purposes only.
6. Proportionate burden means the cost of supplying data should not be excessive and should be assessed in terms of the benefits of the associated statistics.
7. Resources as a principle refers to having sufficient resources to meet the requirements of the code.
8. Frankness and accessibility means that official statistics should be accompanied by information about the quality and reliability of the statistics and it should be accessible to all users.

Three additional protocols and other supplemental texts further outline processes with application to user engagement, release of statistics, and administrative data. Unlike the standards produced by ESS and others, the UK code does not include specific quality standards or indicators. In fact, in describing the principle of sound methods and assured quality, the code explicitly states that “quality should be monitored and assured taking account of internationally agreed practices.”

In September 2013 the Office for National Statistics released version 4.1 of its *Guidelines for Measuring Statistical Output Quality* (Office for National Statistics 2013). Version 4.1 replaces version 3.1, issued six years earlier, and addresses a wider range of statistical data than primarily survey data. In particular, the revised guidelines acknowledge the increasing use of administrative data in the production of statistical output. Defining quality as the familiar “fit for

purpose,” the Guidelines recommend that the producers of statistical output report quality in terms of the five quality dimensions of the ESS.

To assist producers in following this recommendation in reporting on quality, the Guidelines present an extensive set of quality measures/indicators for each quality dimension. The quality measures are organized by the stages of the statistical production process, described as:

- Specifying user needs
- Design and build
- Collection
- Processing
- Analysis
- Dissemination

Each measure/indicator is accompanied by a detailed description and either examples or, if applicable, a formula. In addition, next to the quality dimension that each measure represents is a symbol indicating whether the measure applies only to survey data, only to administrative data, or to either type of data or their combination.

Table II.2 summarizes the distribution of the 131 quality measures/indicators by the five quality dimensions and six stages of the production process. Almost exactly half (65) of the measures represent the accuracy and reliability dimension, and a plurality of measures (53) applies to the analysis stage of production. Accuracy and reliability is the only dimension represented under processing, and it is the most common dimension represented in the design and build, collection, and analysis stages. Relevance is the only dimension represented under specifying user needs while measures representing accessibility and clarity are the most common in the dissemination stage. Seven of the eight measures of timeliness and punctuality occur in the dissemination stage (the other under collection). The 19 measures of accessibility and clarity are

almost evenly split between analysis and dissemination while coherence and comparability are distributed across four of the six production stages.

Table II.1. Distribution of UK quality measures/indicators by quality dimension and stages of the statistical production process

Quality Dimension	Stages of the Statistical Production Process						All Stages
	Specifying User Needs	Design and Build	Collection	Processing	Analysis	Dissemination	
Relevance	8	1			9	4	22
Accuracy and Reliability		7	13	13	28	4	65
Coherence and Comparability		4	4		7	2	17
Timeliness and Punctuality			1			7	8
Accessibility and Clarity					9	10	19
All Dimensions	8	12	18	13	53	27	131

Source: Office for National Statistics (2013).

The quality measures/indicators are too numerous to list in their entirety, but some examples will illustrate their scope. Measures of accuracy and reliability in the analysis stage address not only standard errors and variance calculation generally, but descriptions of methods and models, robustness to model misspecification, analysis of variance to assess the quality of trend estimates, calculation of the “M7” statistic (an indicator of seasonality) and other tests and comparisons relevant to seasonal adjustment, as well as several measures of the impact of statistical disclosure control on output quality. Aspects of statistical disclosure control are also included for the dimensions of relevance, coherence and comparability, and accessibility and clarity. Under dissemination, the measures of timeliness and punctuality deal with various lags while the measures of accessibility and clarity involve either documentation or procedures for obtaining access. The two measures of coherence and comparability within this stage request descriptions of differences between domains and comparison of estimates with other estimates

on the same topic. All four measures of accuracy and reliability under dissemination relate to revisions.

A document from the Bank of England geared toward users of the Bank's data is presented as "...explanatory material describing the relevance of data, how statistics are compiled, the use of imputation methods, and any other information (including quantitative measures) useful to users in their understanding of what the data represent and how they are constructed" (Bank of England 2014). The authors here define quality as "fitness for purpose of published data for users," which is borrowed from the UK Office for National Statistics. This definition of data quality is noted to be purposefully vague to allow for different understandings of quality based on the use of the statistical outputs in question. In addition, the authors describe a framework that they borrowed from the ESS consisting of the same five quality dimensions used in the Office for National Statistics Guidelines discussed above.

4. The Netherlands

Statistics Netherlands drew on several existing quality frameworks to develop its own quality framework, described in *Quality Guidelines 2014: Statistics Netherlands' Quality Assurance Framework at Process Level* (Statistics Netherlands 2014). These included the *European Statistics Code of Practice*, the *Quality Assurance Framework of the ESS*, the data quality assurance framework of the IMF (see below), and a Statistics Netherlands checklist for statistical output (Van Nederpelt 2009). The Statistics Netherlands quality framework and the checklist reflect the application of Object-oriented Quality Management (Van Nederpelt 2010), which was developed at Statistics Netherlands.

The quality framework makes a distinction among the statistical concept, or that which is to be measured; the statistical data, or the estimates of the concept; the statistical output, or how the statistical estimates are presented; and the output release, encompassing what is released and

when. Different dimensions of quality apply to each of these phases, and there are indicators associated with each dimension.

For the statistical concept, four dimensions are relevant. These dimensions and the number of indicators presented for each are:

- Relevance of the statistical concept (7 indicators, such as having documentation that the statistics address known users' needs)
- Coherence of the statistical concept with concepts of other statistics (5 indicators including definitional clarity and uniqueness of variable names)
- Consistency of the statistical concept with reality (1 indicator, reflecting the real world applicability of the statistic as opposed to strictly administrative utility)
- Stability of the statistical concept (1 indicator, reflecting the stability or consistency of the statistic's meaning over time)

For the statistical data, the guidelines address four dimensions as well:

- Accuracy of statistical data (3 indicators involving variance, bias, and their stability over time)
- Comparability of the statistical data (3 indicators involving comparability over time and across subpopulations and their adherence to Eurostat regulations)
- Consistency of statistical data (6 indicators including measures of stability over time and consistency between monthly and annual estimates and with other statistics)
- Confidentiality of the statistical data (1 indicator, reflecting the data's being subject to the appropriate security measures)

For the statistical output there are four dimensions of quality as well:

- Clarity of statistical output (4 indicators including compliance with regulations and the announcement of revisions in advance)
- Accessibility of statistical output (2 indicators reflecting access to internal and external users alike and)
- Completeness of statistical output (5 indicators reflecting coverage of the agreed-upon units and populations, variables, classification system, subpopulations, and reference period)
- Output reproducibility (3 indicators, reflecting minimization of manual adjustments and the application of suitable version control)

For the output release the relevant dimensions are:

- Completeness of the released output (2 indicators reflecting publication in the appropriate web location)
- Timeliness of the release of statistical output (1 indicator reflecting compliance with a Statistics Netherland standard regarding the lag between the reference period of the statistic and its release)
- Predictability of the release of statistical output (2 indicators reflecting compliance with a published schedule or, if not, suitable announcement in advance)
- Punctuality of the release of statistical output (1 indicator reflecting adherence to pre-announced dates over the past 12 months)
- Simultaneous release of statistical outputs (1 indicator reflecting release to all users at the same time and in the same manner)

While several of these indicators are specific to Statistics Netherlands, they can be applied to the production of statistical output by any country.

The checklist is itself an extensive document, which adds several dimensions to those identified in the *European Statistics Code of Practice*. The additional dimensions are:

- Extent of detail—the extent to which subpopulations are distinguished in the statistical output
- Completeness—the extent to which the agreements made with the user on the specifications of the statistic are adhered to
- Numerical consistency—the degree to which the data of different statistics that apply to the same data item equal each other; specific types of coherence are included under this rubric
- Plausibility—the extent to which statistics are “plausible”
- Disputability—the extent to which the accuracy of a statistic may be “opposed” or challenged
- Validity—the extent to which a statistic measures what it is intended to measure
- Reliability—the extent to which a statistic is composed in a reproducible way, although it is noted that reliability is often used in combination with accuracy and that reproducibility is listed separately
- Verifiability—the extent to which the output can be fully retraced from the input data
- Reproducibility—the extent to which statistics have been compiled in a reproducible way, implying fixed algorithms
- Availability—whether statistics continue to be obtainable to users

These additional dimensions, it is noted, are employed “in daily practice” within Statistics Netherlands. As some of the definitions suggest, however, they do not appear to be valued equally. Accordingly, checklist measures are not presented for validity, reliability, verifiability, reproducibility, and availability.

Separate checklists are provided for individual statistics and for parts or all of the statistical program. The checklists consist of measures (whether a certain action has been executed) and indicators (the result of a measuring process, which may be qualitative or quantitative). For example, under relevance one of the measures for an individual statistic is whether agreements have been laid down with the user of the statistic, and one of the indicators is a user’s satisfaction score. For the statistical program a measure of relevance is whether a policy has been formulated on the type of statistics the agency wants to produce or does not want to produce.

We will not review the extensive checklist items, but we do want to highlight what is presented regarding register errors, as these have at least indirect bearing on integrated data. Under the checklist topic “accuracy” there is a subtopic on register errors. Checklist items include:

- Whether audits have been performed on the accuracy of important data items in the register and, if so, whether the results have been described
- Empirical estimates of overcoverage and undercoverage
- What percent of the units and individual data items in the register are not filled (that is, data are missing)
- Several measures of linkability, such as the percentage of records that is linkable, the occurrence of duplicate values among the linking variables, the percentage of linking variables that does not lead to a link, and the percentage of linking variables that leads to an incorrect link

The checklist document also includes a discussion of the relationships among various dimensions of quality. Here it is noted, for example, that the accuracy of a statistic may bear on its relevance in that a statistic that is too inaccurate may lose its relevance to users while

accuracy beyond a certain level may not improve a statistic's relevance. A different type of relationship is found in the trade-off between accuracy and timeliness. To achieve timeliness, it may be necessary to sacrifice some degree of accuracy. This is seen explicitly in statistics that are first released as preliminary, then revised (and possibly revised again). Similarly, accuracy will tend to decline with the extent of detail. Likewise, preserving confidentiality may require a reduction in accuracy. Other relationships include the fact that numerical *inconsistency* can lead to a reduction in plausibility and an increase in disputability.

5. Finland

We include Finland (and Sweden in the next section) to illustrate how the Eurostat quality dimensions are represented in the official documents of European Union members.

Key principles governing the production of statistics in Finland are laid out in the Finnish Statistics Act (http://www.stat.fi/meta/lait/lait_statisticsact04.pdf), which was adopted by Finland's Parliament in 2004. As stated at the end of the Act's first chapter:

The objective of this Act is to ensure the availability of reliable statistical information required in social decision-making and planning and in fulfilling obligations under international statistical co-operation by harmonizing and rationalizing the principles and procedures applied in the collection, processing, use, release and storing of data, to promote the observation of good statistical practice in the National Statistical Service and to ensure that the rights of those who provide data for statistical purposes or whom the data concern are upheld.

The Act covers the types of data that may be collected and the authority of Statistics Finland to collect such data, compile and publish statistics, and release confidential data—subject to restrictions. The Act also defines the rights and obligations of those from whom data are requested. However, the Act does not mention quality—either as a goal or as something to be assessed.

The agency's *Guidelines on Professional Ethics* (Statistics Finland 2006) define six ethical principles to which its employees must adhere:

- Impartiality
- Reliability
- Relevance
- Cost-effectiveness
- Statistical confidentiality
- Transparency

By extension, these principles also apply to the data produced by the agency as well. Hence their resemblance to a number of the quality dimensions discussed above is not surprising.

The agency's *Quality Guidelines for Official Statistics* (Statistics Finland 2007) focus more on production than dissemination. Most of the document is devoted to a step-by-step review of stages in the production of statistics, and both censuses and administrative records and registers are included under the broadly defined term, "statistical surveys."

In its discussion of the publication of statistics, the Guidelines specify that the producers of Official Statistics of Finland (OSF) "must regularly evaluate the quality of the statistics they produce against the quality criteria:"

- Relevance
- Accuracy and reliability
- Timeliness and promptness
- Coherence, consistency and comparability
- Accessibility and clarity

Further, "each set of OSF statistics must be accompanied by a quality report providing a concise assessment of its quality, reliability and applicability for different purposes." A proposed outline of a quality report includes each of the five quality criteria plus a methodological description of the statistical survey.

In addition to preparing the quality report, consideration must be given to the need for a separate and detailed methodological report, which will provide more detail on each of the sections of the quality report plus documentation on any available, archived data files. The precise contents of the methodological report will depend on the nature of the statistics produced and the needs of the end users.

In distributing statistical information, several publishing principles must be observed, which are separate from the requirement to produce a quality report and possible methodological report. These principles—also laid out in the Guidelines—are:

- **Reliability:** All information released shall be accurate and its level of reliability indicated
- **Impartiality:** The information shall be released on schedule and shall be available simultaneously to everybody
- **Immediacy:** All information shall be released as soon as possible after the reference period they describe
- **Clarity:** All information shall be presented clearly, taking into account the needs of end-users. Users of information shall be given every opportunity to draw their own conclusions.
- **Neutrality:** It is important to exercise caution and restraint in the treatment of contentious social issues
- **Interpretation:** All information shall be interpreted and analysed by describing the scale and proportions of different phenomena and by explaining the causes and consequences of changes and phenomena. Where possible, the information contained within a given statistical product shall be compared to other statistical data related to the same phenomenon and to any other relevant information.
- **Timeliness:** All information released shall be tied to current social debate and issues. Statistics Finland shall take the initiative in producing statistical information.
- **Openness:** Reliable statistical information shall not be concealed
- **Guidance:** End-users shall be supported in their acquisition and search for information

In addition, all information must be released simultaneously in Finnish and Swedish, in keeping with the Language Act. Internationally important statistics are released in English as well.

Statistics Finland updated its quality criteria in 2010. The updated criteria, which are posted on the Statistics Finland website (http://www.stat.fi/meta/svt/svt-laatukriteerit_en.html), include the following:

- Impartiality and transparency
- Quality control
- Confidentiality
- Efficiency
- Relevance
- Accuracy and reliability
- Timeliness and punctuality
- Coherence and comparability
- Accessibility and clarity

All nine criteria must be addressed in evaluating the quality of statistics, which presumably means that they must be covered in the quality report that is mandated in the Guidelines.

6. Sweden

Reporting on the quality of official statistics in Sweden is addressed in The Official Statistics Act (2001) and the accompanying Official Statistics Ordinance (2001). A 2013 revision of the Statistics Act introduced the same quality criteria that are included in European legislation (Statistics Sweden 2017). The criteria with their definitions as they appear in the Act are:¹⁷

- **Relevance:** measuring the degree to which statistics meet current and potential needs of the users
- **Accuracy:** the closeness of estimates to the unknown true values
- **Timeliness:** the period between the availability of the information and the event or phenomenon it describes
- **Punctuality:** the time between the date that the statistical agency releases the data and the target date by which the data should be delivered
- **Accessibility and clarity:** the conditions by which users can obtain, use and interpret data

¹⁷ These are a “non-official translation made by Statistics Sweden.”

- **Comparability:** the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas, sectoral domains or over time
- **Coherence:** the adequacy of the data to be reliably combined in different ways and for various uses

The Ordinance includes a section on quality and accessibility, which states that all “statistical agencies shall provide documentation and quality declarations for the official statistics” that they produce.

Under new directives that became effective on September 1, 2016, Statistics Sweden is given a more explicit coordinating role and increased responsibility to follow up on quality throughout the system of official statistics.¹⁸ In response, Statistics Sweden has developed and adopted new regulations that define the agency’s coordinating role. The revised regulations introduce a new quality concept consisting of five main components, which collapse the seven criteria listed earlier by combining the separate criteria of timeliness and punctuality and the separate criteria of comparability and coherence (as does Eurostat). This concept of quality is to be used for all official statistics. Statistics Sweden has also amended an existing regulation that bears on the reporting of quality for official statistics. The amendment includes a new template for quality declarations. A 2016 revision of the Official Statistics Ordinance makes all statistical agencies responsible for evaluating the quality of the official statistics that they produce. On October 31, 2017, Statistics Sweden published two documents that will assist the Swedish statistical agencies in carrying out their new mandate to report on quality: *A Handbook on Quality for Official Statistics of Sweden* and *A Handbook on Evaluation of Quality for Official Statistics of Sweden*. We cannot review their contents, as both were published only in Swedish, and English translations do not yet exist.

¹⁸ Statistics Sweden is one of 27 Swedish statistical agencies as of 2016.

7. OECD

The OECD has produced its own quality framework and guidelines, which are addressed to the statistical activities carried out by the organization (OECD 2011). Defining quality as “‘fitness for use’ in terms of user needs,” the OECD drew on the experience of other statistical organizations in specifying seven dimensions of quality: relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence. The dimension “credibility” is unique to the OECD, which defines this dimension as referring to “the confidence that users place in those (data) products based simply on their image of the data producer, that is, the brand image.” An important aspect of credibility is “trust in the objectivity of the data.” In addition to being influenced by users’ impressions of the producer, such trust is also determined by “the integrity of the production process.”

In presenting its guidelines, the OECD breaks down statistical activities into seven phases:

- Definition of the data requirements in general terms
- Evaluation of other data currently available
- Planning and design of the statistical activity
- Extraction of data and metadata from databases within and external to the OECD
- Implementation of a specific data and metadata collection mechanism
- Data and metadata verification, analysis and evaluation, and
- Data and metadata dissemination

Guidelines are presented for each phase.

We focus on the data and metadata dissemination phase, as transparent reporting of quality applies most directly to this phase. The guidelines include explicit requirements for documentation, noting that the documentation on methodology “must permit users to assess whether the data adequately approximate what they wish to measure and whether data are

produced with tolerances acceptable for their intended use.” The documentation should cover, at a minimum:

- The type of data sources used
- The nature and purpose of the product, as well as the intended uses of the data
- The conceptual universe covered by the data
- Key concepts, variables (or characteristics) and classifications used
- A statement of key accuracy issues, as well as an acknowledgment that the data are subject to error and (if applicable) that the level of error may vary geographically and by other characteristics
- Any variation in accuracy and coherence over time and across countries. The issue of coherence is especially relevant for OECD statistics
- If applicable, a statement advising that the data are subject to revision
- If applicable, a description of benchmarking and seasonal adjustment made to the data and their impact

For statistics that are derived from administrative sources, the OECD asks that several additional topics be addressed in the documentation:

- The purposes for which the data were originally collected
- The merits and shortcomings of the data for the statistical purpose for which they are being used (for example, in terms of conceptual and coverage bias)
- How the data are processed after being received and what, if anything, is done to correct problems in the original dataset
- The reliability of the estimates, including caveats where necessary

These topics would apply, presumably, to estimates derived from integrated data as well as to estimates based entirely on administrative records.

8. IMF

Beginning with an Executive Board discussion in December 1977, interest in the development of a data quality assessment framework for the IMF grew into the drafting of a framework. The current framework, which was published June 25, 2003, incorporates

refinements from an earlier version. We take note of the IMF framework as evidence of how widespread was the production of such frameworks around the turn of the century.

The five quality dimensions included in the *Data Quality Assessment Framework and Data Quality Program* (IMF 2003) are:

- Assurances of integrity—the principle of objectivity in the collection, processing, and dissemination of statistics is firmly adhered to
- Methodological soundness—the methodological basis for the statistics follows internationally accepted standards, guidelines, or good practices
- Accuracy and reliability—source data and statistical techniques are sound and statistical outputs sufficiently portray reality
- Serviceability—statistics, with adequate periodicity and timeliness, are consistent and follow a predictable revisions policy
- Accessibility—data and metadata are easily available and assistance to users is adequate

Each of the five dimensions is represented by one or more elements, and each of these elements is portrayed, in turn, by one or more indicators. For example, the dimension of accuracy and reliability has five elements corresponding to the source data, assessment of the source data, statistical techniques, assessment and validation of intermediate data and statistical outputs, and revision studies. For the assessment and validation element the indicators are:

- 3.4.1 Intermediate results are validated against other information where applicable
- 3.4.2 Statistical discrepancies in intermediate data are assessed and investigated
- 3.4.3 Statistical discrepancies and other potential indicators of problems in statistical outputs are investigated

These indicators reflect recommended practices rather than prescriptions for transparent reporting of quality.

This page has been left blank for double-sided copying.

III. EXTENDING TOTAL SURVEY ERROR TO INTEGRATED DATA

As a framework for describing the sources of error that affect a survey statistic, the TSE model (Groves et al. 2009) has gained wide acceptance and become a valuable tool. Extension of the TSE model to integrated data would seem to provide an equally useful framework for describing and reporting on the sources of error that emerge when different types of data are combined. A particularly notable effort in this direction is found in the work of Li-Chun Zhang of Statistics Norway, who proposed a framework for integrated data based on the TSE model. Adding to our interest in Zhang’s framework, Statistics New Zealand (Stats NZ) has adopted this framework as the basis for its own quality framework for integrated data. In this chapter we review Zhang’s framework in some detail and then discuss Stats NZ’s application of this framework to a statistical system that has given increased emphasis to administrative data use.

A. Zhang’s two-phase framework

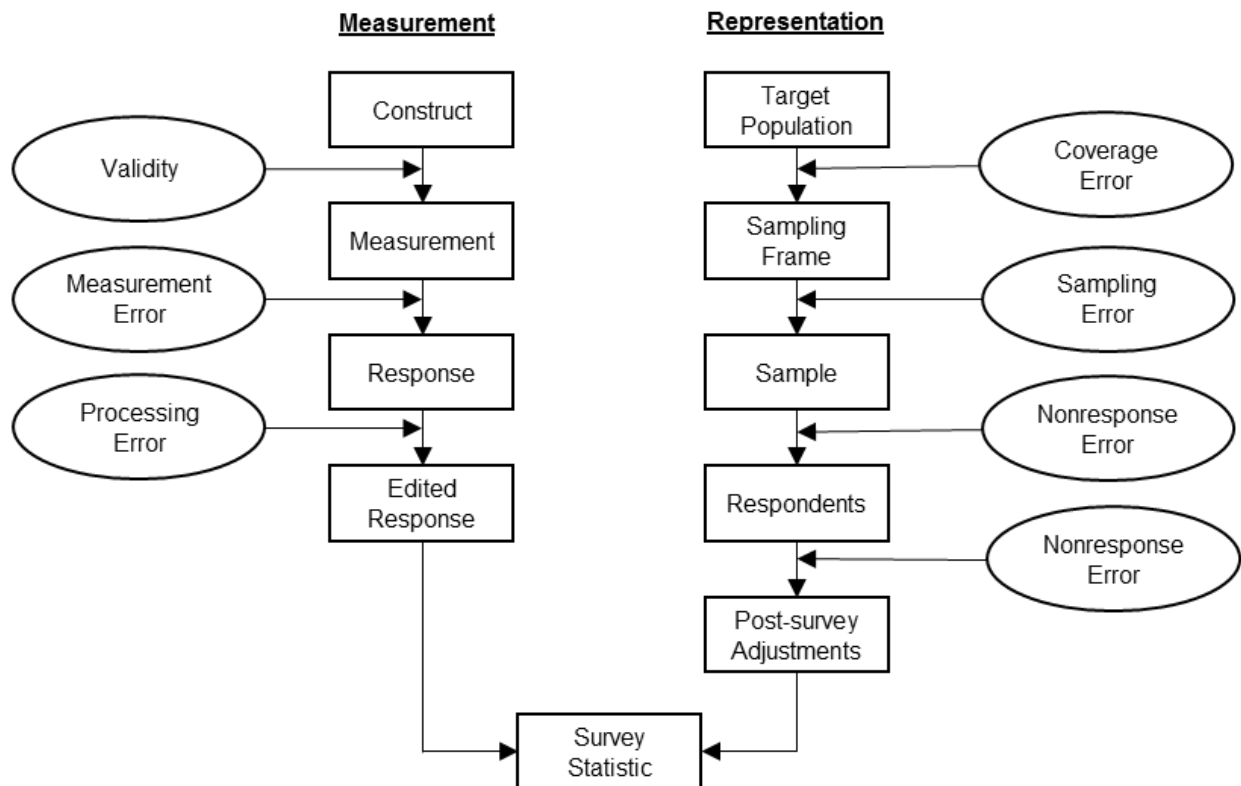
To show how Zhang’s framework builds on this established model, we begin with a brief description of the TSE model.

1. Total Survey Error

The TSE model follows the life cycle of a survey, from conception to the production of a survey statistic. The model builds on the idea that a sample survey consists of a set of questions administered to a sample drawn from a target population. The model traces the dimensions of measurement and representation from their origins in an abstract construct and target population through the survey life cycle—that is, the design and implementation of a sample survey, culminating in a survey statistic. Errors (both random and systematic) may be introduced at each of several stages, which are depicted in the model. In Figure III.1, taken from Figure 2.5 of Groves et al. (2009), the rectangles depict elements of the design of a sample survey, and the ovals are quality concepts that are commonly used with survey data. Each oval describes a

source of error, and the ovals are placed between design elements to indicate that they “reflect mismatches between successive steps” of the survey process.

Figure III.1. Survey life cycle from a quality perspective



Source: Groves et al. (2009).

On the measurement side, the survey life cycle begins with an initial construct and proceeds through operational measurement of the construct, the response of sample members, and the editing of the response. The sources of error are the validity of the measurement followed by the measurement error that occurs between the measurement and the response, and then the processing error that occurs in editing the recorded responses (as well as imputing when responses were not provided). On the representation side, the survey process begins with a target population and proceeds through the construction of a sampling frame, selection of a sample, the participation of respondents, and the application of postsurvey adjustments. The potential sources of error are coverage error due to discrepancies between the sampling frame and the

target population, sampling error arising from the selection of a sample from the sampling frame, nonresponse error due to the failure to obtain responses from a portion of the sample, and adjustment error in the application of the postsurvey adjustments. The processes of measurement and representation culminate in the derivation of a survey statistic from the edited responses of the respondents, adjusted to reflect the target population. Error in the survey statistic is the net result of the errors in measurement and representation depicted in the model.

2. Zhang’s framework

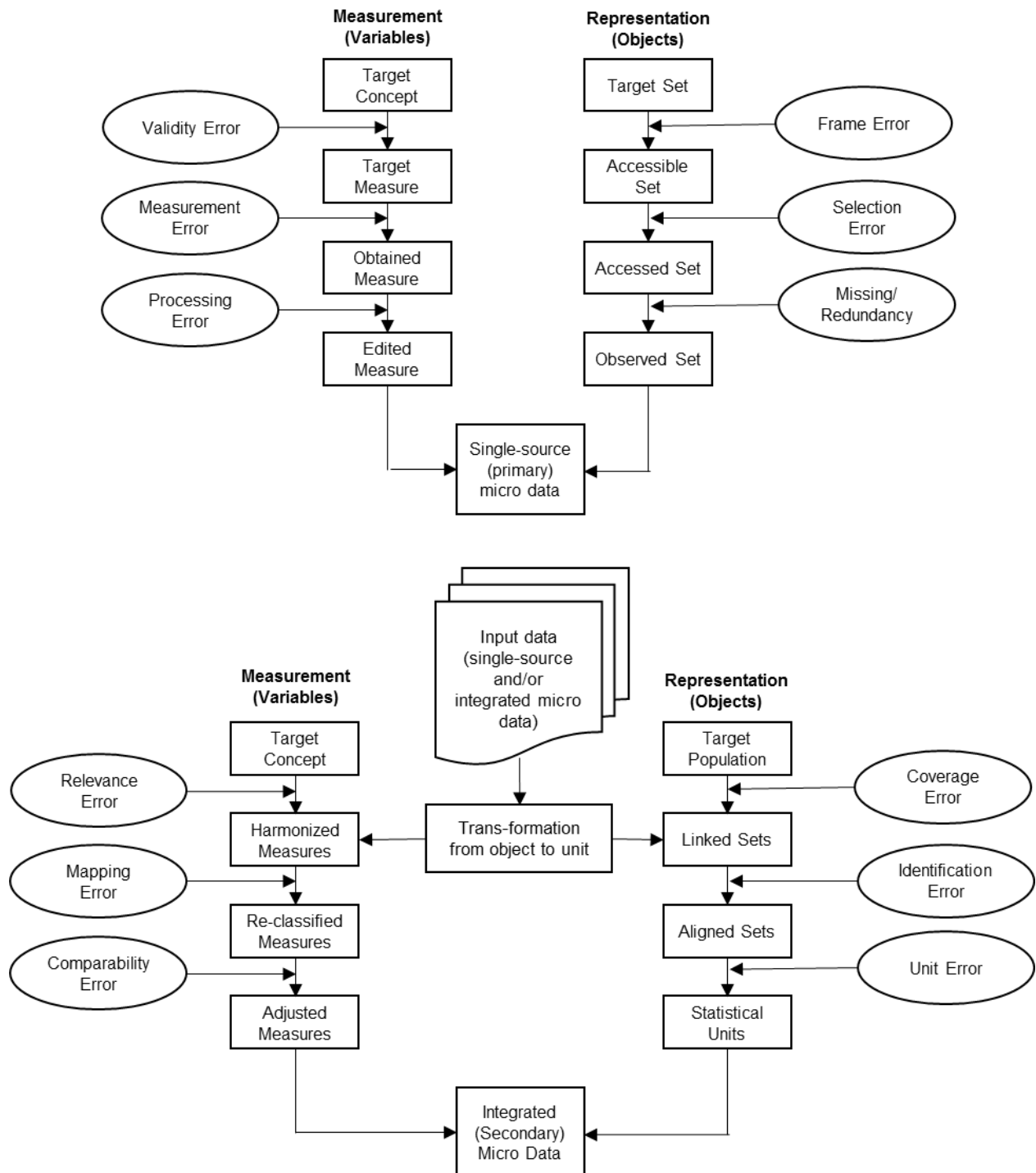
Building on this life-cycle model of potential error sources for sample survey data as well as an adaptation to combined register data by Bakker (2010), Zhang (2012) proposed a “two-phase life-cycle model of integrated statistical micro data,” which is shown in Figure III.2. The first phase describes a single micro data source, but the idea is that each input to the integrated micro data would have its own phase one assessment. Phase two depicts the sources of error characterizing the integrated micro data, where the error components reflect the integration process, which may include transformation of the initial input data. On this latter point, for example, Zhang contrasts the secondary, statistical usage of administrative data with its primary administrative use—that is, the use for which it was created. Zhang observes that administrative data have already gone through a process of conception, collection, and processing prior to any subsequent statistical use. This process is subsumed under phase one.

In motivating the two-phase framework, Zhang observes that administrative register data often have to be combined with data from other sources before they can be used for a statistical purpose, due to limitations in the information collected in the register or how it is organized.¹⁹ In

¹⁹ In the U.S. we have prominent examples of direct uses of administrative data for statistical purposes without the addition of survey or census data—specifically, tax data processed by the IRS and federal Medicare and state Medicaid data compiled by the Centers for Medicare & Medicaid Services.

addition, administrative data may have to be transformed to represent the desired unit (for example, persons instead of transactions). The two-phase framework also recognizes potential secondary uses of survey and census data. Here, too, their primary use would be captured in phase one while the secondary use of the same data would be depicted in phase two, which would account for the (additional) error introduced by the adjustments needed to adapt the original data to their secondary use. Zhang notes a wrinkle with regard to the treatment of census data in phase one, acknowledging that (in the European context), census data themselves may have been generated as integrated statistical data, combining data from multiple registers, perhaps with a survey component as well. In presenting his two-phase framework, Zhang addresses the need for such a framework in his observation that “the 20th century witnessed the birth and maturing of sample surveys; the 21st century will be the age of data integration.”

Figure III.2. Two-phase life cycle of integrated micro data from a quality perspective



Source: Zhang (2012).

Zhang’s phase one is not an exact replication of the TSE model. Most importantly, the end result of phase one—and phase two as well—is a micro dataset, not a single statistic. In addition,

most of the concepts have been renamed to accommodate the inclusion of data from administrative sources, and postsurvey adjustments have been removed from the steps under representation. On the measurement side, the use of target concept, target measure, obtained measure and edited measure is an acknowledgment that the data collected are not necessarily responses to a questionnaire. On the representation side, the use of target set, accessible set, accessed set, and observed set expands the selection mechanism beyond sampling and generalizes the ultimate source of the data beyond respondents. In keeping with this change, nonresponse error in the TSE model is replaced by “missing/redundancy,” which should be read as two terms indicating that some objects of the data collection may be missing while others may be duplicated (redundant). Like the TSE model, however, Zhang’s phase one framework presumes a data collection design that does not readily apply to what Groves and others have termed “organic” or “found” data—references to some types of Big Data. Zhang’s goal is to extend the TSE model to encompass administrative data, recognizing its widespread use in many countries. Other forms of Big Data are off the horizon.

Continuing with the extension of representation to include administrative data, we note that frame error (coverage error) commonly occurs in administrative data when particular activities fall outside the scope of the data collection. Employment data, for example, exclude jobs in the underground economy. This would not be an issue for the primary use of employment data but would become an issue for the secondary use of these data if the target population is broader than persons with employment captured in the administrative system. For administrative sources that capture the entire, applicable universe, selection error may still occur due to events that are not reported or are reported with a delay or are reported with errors that result in their rejection. Lastly, records that reach the final stage of processing may be rejected at that point if, for example, they have too much missing information.

Turning to phase two and its depiction of the integration of multiple data sources, we note, first, that for phase one Zhang uses the term “objects” as the subject of representation versus “units” in phase two. This, he explains, is in recognition of the fact that register data may refer, for example, to jobs, whereas the goal of integration may be data on persons. This transformation of data from object to unit is indicated in the figure with a processing step depicted in a box below the multiple sets of input data.

As with the individual data sources described in phase one, the life cycle of the integrated data begins with a target concept for measurement and a target population (recognizing the shift from objects to units) for representation. On the measurement side, the single target measure of phase one is replaced by multiple harmonized measures in phase two. Zhang describes harmonization as a conceptual alignment that does not involve actually changing the data. Zhang assigns the term relevance error rather than validity error to the discrepancy between the harmonized measures and the target measure because relevance is a term associated with register-based statistics and because relevance is more suitable to the many-to-one relationship that exists between the harmonized measures and the target measure. Changes to the data occur in the next step, which Zhang describes as “turning primary input-source measures into harmonized measures,” and he identifies the error associated with this process as mapping error. The application of editing and imputation yields the final adjusted measures, but Zhang expands the error introduced in this step beyond the processing error in phase one to encompass inherent inconsistency across the data sources, yielding comparability error (or compatibility error, as he uses both terms) if the adjustments are not sufficient to compensate for the inconsistency.

On the representation side, the single accessible set (a generalization of the survey frame) in phase one is replaced by linked sets, whose divergence from the target population is described as coverage error instead of the narrower frame error. The units in the linked sets may not

correspond to the final target units. For example, the linked sets may be defined at the person level, but the units desired in the end may be households. To accomplish this conversion, the persons in the linked sets must be aligned with households, yielding what Zhang terms aligned sets. Errors in this process are defined as identification errors. From the aligned sets the final statistical units are generated. Zhang notes that some of the needed units may not exist in any data source and, therefore, will not be included among the aligned sets. Such units will have to be “created by the statistician” through a process that is invariably imperfect, resulting in unit errors.²⁰

The notion of unit error holds special significance in countries (like Norway) that rely on a population register in conducting their censuses but have no corresponding household register (Zhang 2011). Households must be constructed by assembling the people in the population register into household units based on the information contained in other types of registers or collected in sample surveys. This is a challenging process, and error in constructing these units is a significant concern.

While Figure III.2 does not show this, Zhang’s conceptualization envisions an ideal target integrated dataset. Discrepancies between the target dataset and the final integrated dataset are analogous at the dataset level to the concept of TSE as defined by Groves et al. (2009) at the estimate level. Zhang discusses ways in which the accuracy of the integrated dataset can be assessed. In doing so, he develops the concept of empirical equivalence. Two datasets are empirically equivalent if they generate identical inferences. This does not require that the datasets be identical at the micro level.

²⁰ Unit errors may also result from inaccuracies in the alignment process, which are counted as identification errors.

Zhang extends the concept of empirical equivalence to the assessment of public use data, which deviate from the final integrated data in ways devised to protect the confidentiality of the underlying “true” data. While public use data will differ from the true data, there is an expectation on the part of the user that the public use data should permit very similar inferences (and in many instances identical inferences) as the true data except where restrictions in the information released in the public use data (for example, less detailed geography) clearly limit the inferences that can be generated. Empirical equivalence provides a conceptual basis for assessing the utility of public use data.

Zhang makes one other point that applies to the assessment of integrated data, and this involves validity versus accuracy. In his example, a survey may be designed to provide valid estimates of a concept—for example, the employment rate. With its sample size, the survey will also provide accurate estimates at the national level (that is, characterized by a small mean squared error). Below the national level the survey estimates remain valid (that is, unbiased), but their accuracy declines with the size of the geographic area. An alternative set of estimates based on integrated data may be biased but have no sampling error. While the survey produces more accurate estimates at the national level, the estimates based on integrated data may be more accurate at low levels of geography, where the absence of sampling error gives the integrated estimates smaller mean squared error relative to the survey estimates despite their bias.

B. Stats NZ’s use of the two-phase framework

Zhang’s proposed framework has been adopted by Statistics New Zealand (Stats NZ), which has set a goal of making administrative data its data source of choice, to be “supplemented by survey data collection only when necessary” (Reid, Zabala, and Holmberg 2017). This transformation in Stats NZ’s approach to data collection poses a number of challenges, including how to “assess and explain the quality of statistics that use multiple sources, including

administrative data” (Holmberg and Bycroft 2017). In response to this specific challenge, Stats NZ issued a *Guide to Reporting on Administrative Data Quality* in 2016, which incorporates Zhang’s framework. Consistent with the use of this framework, the Guide covers quality assessments not only for administrative data alone but also for integrated data. The Guide includes quality indicators for each of the phase one and phase two error sources in Zhang’s framework, along with instructions on how to calculate the quantitative indicators. Stats NZ has prepared a metadata worksheet to assist users in compiling the information needed to calculate the phase one quality indicators.

Table III.1 lists the 25 quantitative quality indicators defined by Stats NZ for phase one of the quality framework. Brief descriptions taken from Reid et al. (2017) are included. More extensive descriptions and calculation instructions are provided in Stats NZ (2016). Given that the phase one assessment focuses on the original purpose of the data collection, and not its use in an integrated dataset, these indicators address the original purpose of the data as well. Not all of these indicators will apply to every dataset. Some are clearly appropriate only for survey data and others for only administrative data.

Table III.1. Stats NZ's quantitative quality indicators for phase one

Error source and indicator	
Measurement dimension	
<i>Validity error</i>	
1	Percent of items that deviate from target concept definition
2	Percent of items that deviate from StatsNZ/international standards or definitions
3	Percent of inconsistent records
4	Percent of items affected by respondent comprehension of questions asked in collection process
<i>Measurement error</i>	
5	Item nonresponse rate
6	Item imputation rate
7	Percentage of records from proxies
8	Lagged time between reference period and receipt of data
9	Punctuality
10	Overall time lag
11	Percent of units in administrative data which fail checks
12	Stability of variables
<i>Processing error</i>	
13	Percentage of units of a variable with transcription errors
14	Modification rate--frequency of editing changes to a variable
15	Readability
Representation dimension	
<i>Frame error</i>	
16	Lag in updating population changes--delays in registration
17	Undercoverage--units in the target population not in the accessible set
18	Overcoverage--units in the accessible set not in the target population
19	Authenticity--correctness of identifiers
<i>Selection error</i>	
20	Adherence to reporting period
21	Dynamics of births and deaths--changes in rates over time
22	Inconsistent objects/units
<i>Missing/redundancy error</i>	
23	Unit nonresponse rate
24	Percentage of duplicate records
25	Percentage of units that have to be adjusted to create statistical units

Source: Statistics New Zealand (2016).

For phase one Stats NZ has also defined a number of qualitative indicators of quality. Most of these indicators ask for descriptions of aspects of the collection and processing of an input dataset. Table III.2 lists qualitative indicators for the measurement dimension, and Table III.3 lists qualitative indicators for the representation dimension.

Table III.2. Qualitative quality indicators for phase one measurement

Error source and indicator
Validity
Describe the primary purpose of the data collection for each source
Describe the main uses of the administrative dataset
Describe differences in concepts, definitions, and classifications
Describe the data collection method
Describe the reference period for the data collection
Describe changes over time in the administration of data collection and assess the likely impact of these on the definition of concepts and classifications
Measurement error
Describe processes employed by the administrative data to reduce measurement error
Context bias
Noise/seasonal variation
Rounding error and rounding/heaping
Detecting missing values
Imputation methods
Processing error
Describe the main sources of processing error
Describe the data processing known to be required on the administrative data source in terms of the types of edits carried out
Describe the data processing known to be required on the administrative data source to deal with nonresponse
Quality control
Skill level of coders/editors
System bias
Use of standard classifications
Extent of data manipulation
Confidentialization method
System changes

Source: Statistics New Zealand (2016).

Table III.3. Qualitative quality indicators for phase one representation

Error source and indicator
Frame error
Describe the common identifiers of population units in the administrative data
Mapping of reporting units to statistical units
Population definition
Changes in population coverage
Duplicates
Updating of reporting units
Describe the extent of coverage of the administrative data and any known coverage problems
Describe methods used to deal with coverage issues
Selection error
Describe any issues with classification and how these issues are dealt with
Missing/redundancy error
Detecting duplicate records
Methods of treating duplicate records
Describe differences between responders and non-responders

Source: Statistics New Zealand (2016).

Turning to phase two, Table III.4 lists the 19 quantitative indicators that Stats NZ has defined for the reporting of quality for integrated data, beginning this time with the representation dimension. There are as yet no qualitative indicators for phase two.

Table III.4. Stats NZ's quantitative quality indicators for phase two

Error source and indicator	
Representation dimension	
<i>Coverage error</i>	
1	Undercoverage--proportion of units in the target population missing from the final dataset
2	Overcoverage--proportion of units in the final dataset not in the target population
3	Proportion of units linked from each dataset to a base dataset, or percentage link rates between pairs of datasets
4	Proportion of duplicated records in the linked data
5	False positive and negative link rates
6	Macro-level comparisons of the distribution of linked objects with reference distributions
7	Delay in reporting--time lag between end of reference period and receipt of final data
8	Linking methodology used
<i>Identification error</i>	
9	Proportion of units with conflicting information
10	Proportion of units with mixed or predominance-based classifications
11	Rates of unit change from period to period
<i>Unit error</i>	
12	Proportion of units that may belong to more than one composite unit
Measurement dimension	
<i>Relevance error</i>	
13	Percentage of items that deviate from Statistics NZ/international standards or definitions
<i>Mapping error</i>	
14	Proportion of items that require reclassification or mapping
15	Proportion of units that cannot be clearly classified or mapped
16	Distribution of variables in linked data
17	Indicators and measures of modeling error
<i>Comparability error</i>	
18	Proportion of units failing edit checks
19	Proportion of units with imputed values

Source: Statistics New Zealand (2016).

Reid et al. (2017) add a third phase to Zhang's framework that provides for assessing the quality of final outputs—that is, the statistical estimates derived from the integrated micro dataset that is the endpoint of phase two. These estimates may incorporate a variety of statistical or econometric techniques, ranging from simple summations to the application of complex models and may also include additional adjustments—for example, for seasonality. The third

phase also takes account of the inaccuracies arising from efforts to compensate for the errors introduced in phases one and two. Unlike the quality indicators that Stats NZ has developed for phases one and two, however, standard quality indicators for phase three do not yet exist.

The methods that may be applied in phase three are varied, and they must be tailored to address each unique application. Underscoring this point, Reid et al. presented three case studies that were used to test and further develop the three-phase framework. The case studies demonstrate three different approaches to evaluating final estimates derived from integrated data. The first case study involved a redesign of the Building Activity Survey, in which a sample survey component was to be replaced with modeled values derived from administrative data on building consents (analogous to building permits in the U.S.). The second case study involved the prospective replacement of personal income measures in a household survey with data obtained from linked tax records. The third case study involved consideration of an approach to population estimation based on imperfectly linked administrative sources.

In describing the first case study, we focus on the modeling component, which the authors describe as being applicable to any situation where the responses to a particular variable in a survey can be approximated by applying a statistical model to a closely related variable or variables from administrative data. With the redesign, only large construction jobs would continue to be surveyed while estimates modeled with administrative data would replace the survey responses for smaller (non-large) jobs. For the phase two evaluation, where to locate the modeling error was an issue. In effect, the responses that were not collected were imputed, but because the responses were missing by design rather than item nonresponse, Stats NZ was reluctant to treat the modeling error as imputation error. Instead, the modeling was interpreted as converting the administrative data into a harmonized measure, making the modeling error more akin to mapping error. Not all modeling error can be treated the same way, however. Part of the

motivation for adding a third phase to Zhang’s framework was to accommodate the processes involved in taking the phase two unit record data as an input and applying statistical techniques to generate final outputs. In reporting each value from the redesigned survey, Stats NZ provides an estimate of modeling error, the proportion of the value that was modeled rather than surveyed, and the imputation rate.

The second case study is just one example from a long-term evaluation of the prospects of replacing components of the New Zealand census with administrative records. Central to this effort is Stats NZ’s Integrated Data Infrastructure (IDI), a compilation of multiple administrative datasets from several government agencies. A component of the IDI is a list of individuals created from the union of tax, birth, and long-term visa records. The target population for this list, which is called the “spine,” is all persons who have ever resided in New Zealand. The spine serves as a central database to which all other datasets can be linked. As part of its long-term census research, Stats NZ has linked records from the 2013 Census to the spine. This makes it possible to explore potential replacement of census questions with a variety of elements extracted from administrative records.

The New Zealand census collects sources of income and, for each source, an income range. In addition to nonresponse, the census income data can be affected by a variety of types of reporting error. Substitution of income data from the tax system for data that would otherwise be collected in the census can be done for persons whose census records and tax records both link to the spine. In the prototype test, around 94 percent of the census records could be linked to the spine. The false positive rate was estimated at 0.7 percent. Linkage errors are due primarily to low quality linking information in the census (names and dates of birth being the main factors). The tax data are linked to the spine using unique tax identifiers, so linkage issues are not significant. However, persons may be represented in the tax data who were not included in the

census, and persons included in the census may be absent from the tax data. The non-matches between the census and the spine are a source of uncertainty with respect to tax records that do not link to census records.

The critical question in the phase three evaluation is whether the conceptual mismatch between the tax data—which does not include all of the sources of income included in the census—and the target measure of income (gross total income) is a source of greater error than the reporting error in the census income data. The conceptual mismatch is regarded as a phase two error, which arises from using the tax data for a different purpose than the one for which they were first collected. The phase one error for the tax data, which reflected its original purpose, may have been negligible. The measurement error in the census data is a phase one error. In its evaluation, Stats NZ found that despite the known exclusions from the income captured in the tax data, the amounts obtained from the matched administrative records were generally higher than those reported in the census.

Like the second case study, the third case study also derives from the long-term effort to redesign the New Zealand census. With the data held in the IDI, it may be possible to estimate the size of the population directly. Attempting to do so will shed light on the limitations of administrative data generally and on the strengths and weaknesses of individual sources, which may suggest potential improvements. In the phase two evaluation, the major sources of error lie on the representation rather than measurement side. Coverage error in both directions is likely, as is linkage error. Because linkage plays a role in unduplicating the multiple administrative sources, false negative links (essentially, failure to recognize that two records from different sources represent the same person) result in duplicates in the estimated population. The final integrated dataset that represents the end point of phase two will have significant over- and undercoverage. The need in phase three is for an estimation procedure that will correct for these

errors. The problem can be characterized as one of constructing a model that will describe who in the administrative data ends up in the final dataset and who in the target population is not included or represented in any of the administrative datasets. Such a model is still a work in progress, but by helping to understand the sources of error and their impact, evaluating such a model through the three-stage quality framework can provide a path of continuing improvement.

This page has been left blank for double-sided copying.

IV. QUALITY ASSESSMENT IN THE USE OF ADMINISTRATIVE DATA FOR OFFICIAL STATISTICS

Distinct from the literature on quality frameworks and quality reporting in general is a literature focusing on quality issues in the application of administrative records to the production of official statistics—that is, the statistics generated by national statistical organizations in the performance of their defining functions. This literature includes work generated by international organizations like the U.N. and the European Union, national statistical offices, and individual researchers—often affiliated with national statistical offices. Some of this literature addresses questions related to transparency in reporting of quality but without using those terms. Here we examine a selection of works that speak to the issues that are most central to our review. This literature is drawn from the UNECE, the European Commission, the UK Statistics Authority, and Statistics Netherlands.

A. The UNECE

In 2011 the UNECE released *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices* (UNECE 2011). In the Foreword the authors comment on the need for such a document:

Although several subject specific texts exist, there have, until now, been no general, international methodological guidelines to help those in the early stages of using administrative data. This handbook aims to fill that gap. It builds on material developed over ten years in the context of an international training course on the use of administrative sources for statistical purposes. That course has now been delivered over ten times, to audiences of official statisticians from throughout Europe, Western and Central Asia, and North Africa.

While most of the content focuses on ways to use administrative sources and the issues that must be addressed in doing so, and does so at a fairly elementary level, the Handbook also includes a chapter on quality, which we summarize below. With regard to integrated data, there are chapters on data linkage and matching and on using administrative data to supplement

statistical surveys, but neither of these chapters addresses quality assessment of the integrated data or the integration process.

In discussing quality, the Handbook cites agreement among the major international agencies on the following criteria for evaluating the quality of statistical data:

- Relevance—the degree to which statistics meet the needs of current and potential users
- Accuracy—the closeness of statistical estimates to true values
- Timeliness—the length of time between data being made available and the event or phenomena they describe
- Punctuality—the time lag between the date that data were actually released and the target release date
- Accessibility—the physical conditions in which users can obtain data, including the forms and format
- Clarity/interpretability—whether data are accompanied by sufficient and appropriate metadata, including graphs and maps, and whether information on quality is available
- Coherence/consistency—whether data from different sources convey the same message to users
- Comparability—the extent to which differences between statistics can be attributed to differences between the true values versus methodological differences; comparability can be defined over time, over countries or regions, and between statistical domains

With regard to relevance, the Handbook adds that this dimension refers to whether the statistics that are needed are produced and whether the statistics that are produced are needed.

The Handbook notes that these criteria can be used to assess the quality of the statistics that are the end result of the use of administrative records or to evaluate the quality of different administrative sources prior to their use. When the quality of a potential administrative source is being evaluated, accuracy may be difficult to assess in the absence of sufficient information on the population covered by the data and the process of collecting the data. In this case, the Handbook recommends that consideration be given to the credibility of the data source and the plausibility of the data when compared to other sources. When sufficient information is available, what is important to assess is how well the administrative units and variables

approximate the units and variables needed for statistical purposes. The Handbook defines such an assessment as bearing on the quality criterion of coherence.

While cost considerations are generally viewed as a constraint in the collection of data, the low cost of administrative data relative to survey data collection may allow acceptance of lower quality in some dimensions in choosing an administrative source over a more costly alternative source. It is possible as well that a portion of the cost savings can be used to finance enhancements to the quality of the administrative data.

A final few points on quality are, first, that an assessment for each of the three stages of input, data processing, and output is essential and, second, that the availability of good metadata at each stage is vital to such an assessment. In addition, the Handbook notes that the views of users are critical in evaluating the quality of statistical outputs.

B. The European Commission

The European Commission has funded research on a range of topics related to administrative data use for official statistics. One series of projects is being undertaken by BLUE-Enterprise and Trade Statistics (ETS) and involves a series of “work packages” with different themes.²¹ For example, Work Package 4 is dedicated to improving the use of administrative sources. In a paper prepared under this package, Laitila et al. (2011) discuss alternative ways of using administrative data in the preparation of official statistics and how to assess the quality of such data at the stages of input, production, and output. They present a number of indicators, many of which are similar to those discussed earlier. For example, when an administrative source is integrated with a base register (one way in which an administrative source may be used), mismatches between the two may reflect under-coverage in the base

²¹ BLUE-ETS is coordinated out of the Italian National Institute of Statistics (ISTAT).

register, under-coverage in the administrative source, overcoverage in the base register, or over-coverage in the administrative source. Indicators are defined for all four possibilities.

While the user's focus is on the quality of statistical outputs, the producer must be concerned with input quality and with production process quality. A major focus of Work Package 4 is the development of a standardized way—using an instrument—to assess the suitability of an administrative data source as a potential input to a statistical process. In another paper prepared under this work package, Daas and Ossen (2011) distinguish explicitly between the quality of the source for its original purpose, which they term Data Source Quality, and the quality of the source for its specific statistical purpose, which they label Input Oriented Output Quality.

Daas et al. (2011) propose a list of quality indicators for administrative data when used as an input data source for national statistics. The indicators are grouped under five general dimensions of quality: technical checks, accuracy, completeness, integrability, and time-related factors. The addition of a dimension of technical checks to the more familiar dimensions of accuracy, completeness, coherence, and time-related considerations acknowledges that a dataset must satisfy certain technical requirements to be usable. Similarly, the dimension of integrability recognizes that an administrative data source will often be combined with one or more other data sources when used to produce official statistics. In developing their initial indicators, the authors made other adjustments to the quality dimensions. Coherence was divided into two components: coherence within the dataset and coherence between datasets. Internal coherence was then incorporated into the accuracy dimension. Indicators of stability were added to the time dimension in recognition of the importance of a new issuance of a dataset resembling previous issuances.

The indicators, which are presented in Table IV.1, draw on the first phase of Zhang’s (2012) two-phase model, discussed in depth in the preceding chapter. Following Zhang, there are indicators that correspond to objects (the representation side of Zhang’s diagram) and indicators that correspond to variables (the measurement side). The dimension of integrability bears most directly on the integration of multiple sources. The four indicators, which bear on different aspects of record linkage, are intended to capture how well the data source can be integrated into the statistical production system of an organization.

Table IV.1. Quality indicators for administrative data used as an input source

Dimension	Indicators	Description
1. Technical checks		Technical usability of the file and data in the file
1.1	Readability	Accessibility of the file and data in the file
1.2	File declaration compliance	Compliance of the data in the file to the metadata agreements
1.3	Convertability	Conversion of the file to the organization’s standard format
2. Accuracy		Closeness of the objects and variables to the exact/true objects and values defined and the extent to which data are correct, reliable, and certified
<i>Objects</i>		
2.1	Authenticity	Legitimacy of objects
2.2	Consistency	Overall consistency of objects in source
2.3	Dubious objects	Presence of untrustworthy objects
<i>Variables</i>		
2.4	Measurement error	Deviation of actual data value from ideal error-free measurements
2.5	Inconsistent values	Extent of inconsistent combinations of variable values
2.6	Dubious values	Presence of inconsistent combinations of values for variables
3. Completeness		Degree to which a data source includes data describing the corresponding set of real-world objects and variables
<i>Objects</i>		
3.1	Undercoverage	Absence of target objects (missing objects) in the source (or in the business register)
3.2	Overcoverage	Presence of non-target objects in the source (or in the business register)
3.3	Selectivity	Statistical coverage and representativity of objects (incomplete coverage of target population)
3.4	Redundancy	Presence of multiple registrations of objects
<i>Variables</i>		
3.5	Missing values	Absence of values for (key) variables

Dimension	Indicators	Description
3.6	Imputed values	Presence of values resulting from imputation actions by administrative data holder
4. Time-related dimension		Indicators that are time and/or stability related
4.1	Timeliness	Lapse of time between the end of the reference period and the moment of receipt of the data source
4.2	Punctuality	Possible time lag between the actual delivery date of the source and the date it should have been delivered
4.3	Overall time lag	Overall time difference between the end of the reference period in the source and the moment the organization has concluded that it can definitely be used
4.4	Delay	Extent of delays in registration
<i>Objects</i>		
4.5	Dynamics of objects	Changes in the population of objects (new and dead objects) over time
<i>Variables</i>		
4.6	Stability of variables	Changes of variables or values over time
5. Integrability		Extent to which the data source is capable of undergoing integration or of being integrated
<i>Objects</i>		
5.1	Comparability of objects	Similarity of objects in source--at the proper level of detail--with the objects used by the organization
5.2	Alignment	Linking-ability (align-ability) of objects in source with those of the organization
<i>Variables</i>		
5.3	Linking variable	Usefulness of linking variables (keys) in source
5.4	Comparability of variables	Proximity (closeness) of variables between the source and similar variables in other sources used by the organization

Source: Daas et al. (2011).

In a follow-on to Daas et al. (2011), Daas and Ossen (2011) proposed mostly quantitative measures for each indicator. For example, the measure of undercoverage is the percent of objects from the reference population missing from the source data, and the measure of overcoverage is the percent of objects in the source data that are not included in the reference population. Both of these measures presume that a list of members of the reference population exists. If no such list exists, then the organization must first produce such a list. To measure selectivity, the most rigorous suggestion is to calculate a Representativity indicator (Schouten et al. 2009), which captures differential representation by stratum. As with the measures of undercoverage and

overcoverage, however, the ability to calculate a Representativity indicator is contingent on the existence of data on a reference population with which the source data may be compared.

C. UK Statistics Authority

In July 2014 the UK Statistics Authority published a draft report for comment (an “exposure draft”) entitled, “Quality Assurance and Audit Arrangements for Administrative Data” (UK Statistics Authority 2014). This was followed in January 2015 by the issuance of a regulatory standard in the form of two brief documents, “Quality Assurance of Administrative Data: Setting the Standard” (UK Statistics Authority 2015a) and “Administrative Data Quality Assurance Toolkit” (UK Statistics Authority 2015b). The draft report stands out among the administrative data quality assurance literature in expressing both optimism and concern about the prospects of greater use of administrative data in official statistics. In fact, the report gives more attention to the challenges than the benefits of using administrative data for statistical purposes. We review some of the concerns expressed in the draft report and then summarize the recommendations incorporated into the two documents that followed.

While acknowledging that administrative data can be an important source for official statistics, the draft report finds that there is a risk that the producers of official statistics may assume without justification that administrative data are more reliable than survey-based data. While survey data are often subject to quality checks at each stage of collection and processing, this may not be true of administrative data. Likewise, while uncertainty and bias are acknowledged as concerns with survey data, and effort is expended to reduce their impact on the final estimates, this is less common with administrative data. Case studies presented in the report highlight good practices, but the authors conclude that “the focus of the quality assurance of administrative data needs to be widened to encompass critical thinking about the entire statistical process, including the data recording and collection stages” (UK Statistics Authority 2014).

To address these concerns, the report proposes the use of a quality assurance matrix that is presented in the report and offers guidance in the form of questions that should be asked about the statistics and their producers by non-statisticians who use official statistics based on administrative data. We focus our attention on the matrix.

The quality assurance matrix presented in draft form in July 2014 and in final form in January 2015 (in the toolkit document) includes four practice areas:

- The operational context and administrative data collection
- Communication with data suppliers
- Suppliers' quality assurance principles, standards and quality checks
- The producer's quality assurance investigations and documentation

It can be seen that each of these practice areas is focused on the quality of the administrative data as an input to the statistical production process.

For each of these four practice areas the matrix lists actions or activities corresponding to four levels of quality assurance: no assurance, basic assurance, enhanced assurance, and comprehensive assurance. Basic assurance implies that the statistical producer reviews the administrative data QA arrangements and publishes a high-level summary of the assurance. Enhanced assurance means that the statistical producer evaluates the administrative data QA arrangements and publishes a fuller description of the assurance. Comprehensive assurance indicates that the statistical producer investigates the administrative data QA arrangements and the results of an independent audit and publishes detailed documentation about the assurance and audit. The toolkit includes a risk/profile matrix that can be used to determine the level of quality assurance that is needed or appropriate given the likelihood that quality issues may arise in the data and the importance of the statistics that will be produced from the data.

For the fourth practice area, for example, the producer's quality assurance investigations and documentation, the actions that would constitute a comprehensive assurance are the following:

- Provide a detailed description of own quality assurance checks on the administrative data
- Give quantitative (and, where appropriate, qualitative) findings for specific quality indicators
- Undertake comparisons with other relevant data sources (such as survey or other administrative data)
- Identify possible distortive effects on targets
- Identify the strengths and limitations of the administrative data and any constraints on use for producing statistics
- Explain the likely degree of risk to the quality of the administrative data provided by the operational context and data collection approach

Again, the goal of these activities is to ensure that the quality of the administrative data suits the data's intended use.

D. Statistics Netherlands' quality framework for administrative data

Statistics Netherlands has developed a quality framework explicitly for administrative data with the goal of being able to assess the quality of an administrative data source in an efficient and standardized way (Daas et al. 2009). The framework provides for a multi-level view of the quality of a data source. At the highest level are three hyperdimensions: (1) Source, (2) Metadata, and (3) Data. The three hyperdimensions describe different aspects of the quality of a data source, and these affect the usability of a data source in different ways. Also, the three hyperdimensions are ordered from the most general (Source) to the most detailed (Data). Below each hyperdimension is a set of dimensions, which differ across the hyperdimensions. Associated with each dimension is a set of quality indicators, each of which is measured or estimated by one or more methods, which can be qualitative or quantitative.

We refer the reader to Daas et al. (2009) for a complete listing of the indicators and measures under each hyperdimension, but to convey a sense of what differentiates this quality

framework from others described in this report, we review the dimensions under each hyperdimension. The Source hyperdimension encompasses five dimensions: (1) supplier, (2) relevance, (3) privacy and security, (4) delivery, and (5) procedures. Procedures refer to such things as how the data are collected, planned changes in the data source, how to contact the data source keeper in the event of problems, and the steps to be taken if the data are not delivered as arranged. The Metadata hyperdimension contains four dimensions: (1) clarity, (2) comparability, (3) unique keys, and (4) data treatment by the data source keeper. Unique keys here refers to the presence of identification keys and unique combinations of variables. The Data hyperdimension has 10 dimensions: (1) technical checks, (2) overcoverage, (3) undercoverage, (4) linkability, (5) unit nonresponse, (6) item nonresponse, (7) measurement, (8), processing, (9), precision, and (10) sensitivity. Processing refers to editing, imputation, and outlier correction. Sensitivity includes such indicators as the frequency of missing values, the selectivity of the composition of the dataset (for example, as measured by an R-index), and the effects of these on totals (as measured by their maximum bias, for example).

The authors have developed a checklist for evaluating the Source and Metadata hyperdimensions.²² The Data hyperdimension does not lend itself to a checklist approach because the measures that are used for evaluation require extensive calculations. For the other two hyperdimensions, though, the checklist can be completed in a short amount of time. If they suggest problems with a data source, there may be little point in investing a more significant amount of time in evaluating the Data hyperdimension.

²² With respect to its purpose and some aspects of its design, the checklist resembles the *Data Quality Assessment Tool for Administrative Data* (Iwig et al. 2013), which was prepared to help users in the U.S. assess the fitness of an administrative data source for an alternative, statistical use. The 43 questions included in the Tool request more information than most of the checklist items and focus less on applications to official statistics, but both the Tool and the checklist are intended to provide relatively quick assessments of an administrative data source prior to use (or, for the checklist, prior to a more extensive evaluation).

The checklist is reproduced in the appendix to Daas et al. (2009), and the use of the checklist to evaluate six administrative data sources is illustrated. A review of some findings from this evaluation will show the types of information that the checklist is able to provide. One of the data sources scores poorly on the delivery dimension because it is rarely delivered on time. The same data source also scores poorly on the clarity and comparability dimensions of the metadata hyperdimension—mostly because of a discrepancy between the definition of a key variable in the data source and the definition used by Statistics Netherlands. The data treatment dimension proved difficult to assess for five of the six data sources due to the very limited information that Statistics Netherlands was able to obtain on the checks and modifications performed by the keeper of each data source.

This page has been left blank for double-sided copying.

V. BIG DATA AND OFFICIAL STATISTICS

There is a growing international literature on the use of Big Data in official statistics, and even the recent American Association for Public Opinion Research (AAPOR) Big Data Task Force had an international membership (AAPOR Big Data Task Force 2015). Australia, Italy, and the Netherlands are among the countries whose statistical organizations have launched Big Data programs (Tam and Clarke 2015, Daas et al. 2015). Unlike the situation with administrative records, the nations of Europe and other parts of the world do not hold an advantage over the U.S. in terms of their prior experience with Big Data. Nevertheless, international efforts to establish the usefulness of Big Data as a source for official statistics are notable. Regular conferences are devoted to the topic, and the UN has established a Global Working Group on Big Data for Official Statistics. This chapter focuses on the Working Group's efforts to develop a quality framework for Big Data. The chapter concludes with brief discussions of work by an IMF Internal Group on Big Data and the AAPOR Task Force.

A. UN Global Working Group on Big Data

Created in March 2014, the UN Global Working Group on Big Data for Official Statistics (<https://unstats.un.org/bigdata/>) includes 22 member countries and nine international organizations. As stated on its website, the Working Group's goals are to "adequately address issues pertaining to methodology, quality, technology, data access, legislation, privacy, management and finance, and provide adequate cost-benefit analyses on the use of Big Data." The members collaborate on and share findings from pilot studies, feasibility assessments, and exploratory research on the use of Big Data for official statistics. While the Working Group has yet to produce a set of official standards, it has made progress in developing a Big Data quality framework.

In December 2014, a UNECE Big Data Quality Task Team published *A Suggested Framework for the Quality of Big Data*. This effort, which is separate from but cited in a report of the Working Group (UN Economic and Social Council 2015), grew out of an April 2013 meeting of the UNECE Expert Group on the Management of Statistical Information Systems, which named Big Data a challenge for official statistics. A proposal was developed, and a project on “The Role of Big Data in the Modernisation of Statistical Production” was undertaken the following year. Four task teams were established to address different aspects of the problem. These teams were the Privacy Task Team, the Partnerships Task Team, the Sandbox Task Team, and the Quality Task Team. The Quality Task Team included representatives of the national statistical offices of Australia, Canada, France, Italy, Mexico, Poland, and Slovenia as well as the Statistical Division of the UN.

The quality team studied existing quality frameworks designed for survey and administrative data but concluded that “the application of either traditional data quality frameworks or those designed for administrative data would be an inadequate response to Big Data.” Frameworks designed for administrative data tended to have a broader scope and an ability to deal with a wider variety of data sources and data types than frameworks designed for survey data, but the scope of Big Data exceeds that of administrative data, requiring a different approach.

Three general principles underlay the development of the proposed framework. The first is that “fitness for use” remains a central principle in assessing the quality of a data source. The second is that the framework should be generic and flexible and able to apply its quality dimensions to the three phases of input, throughout, and output. The third is that the framework allow an assessment of effort versus gain—that is, a determination of whether the effort involved in obtaining and analyzing the data is worth the benefits gained from doing so.

The quality framework includes the following dimensions:

- Institutional/business environment—the organizational factors that may have a significant influence on the effectiveness and credibility of the agency producing the data
- Privacy and security—the institutional and organizational factors for both the data provider and the data producer that may have a significant influence on the intended use of the data, given legal limitations, organizational restrictions, and confidentiality and privacy concerns
- Complexity—the lack of simplicity and uniformity in the data
- Completeness—the extent to which metadata are available to afford a proper understanding and use of the data
- Usability—the extent to which the statistical organization will be able to work with the data without the need for specialized resources or the imposition of an excessive burden
- Time factors—the timeliness and periodicity of the data
- Accuracy—the degree to which the information correctly describes the phenomena it was designed to measure; a key concern with respect to Big Data is selectivity or its lack of representativeness
- Coherence—the extent to which the dataset follows standard conventions, is internally consistent, consistent over time, and consistent with other data sources; another key aspect of coherence is linkability, or the ability to be linked or merged with other relevant datasets
- Validity—the extent to which the dataset measures what the user is attempting to measure
- Accessibility and clarity—the ease of access to the data and metadata and the availability of unambiguous descriptions
- Relevance—how well the statistical product meets the needs of users in terms of the concept(s) measured and population(s) represented

Drawing on the Statistics Netherlands' quality framework for administrative data, discussed in the preceding chapter, these dimensions are nested within the three hyperdimensions of Source, Metadata, and Data. The nature of the nesting varies across the input, throughput, and output phases.

The input phase includes those activities associated with the initial acquisition of the data. These activities may encompass assessing the suitability of acquiring a dataset and assessing the quality of the dataset once it has been acquired. Table V.1 shows the interrelationships among the hyperdimensions and the dimensions during the input phase. During this phase the source hyperdimension has two quality dimensions: the institutional/business environment and privacy and security. The metadata hyperdimension has seven quality dimensions, and the data

hyperdimension has four. Some of the quality dimensions—specifically coherence-linkability, coherence-consistency, and validity—appear under both the metadata and data hyperdimensions. For each dimension the table also lists one or more factors to consider. For example, for the complexity dimension under the metadata hyperdimension, the factors to consider include technical constraints, whether the data are structured or unstructured, the readability of the data, and the presence of hierarchies and nesting.

Table V.1. Dimensional structure of the input phase of the UNECE Big Data quality framework

Hyperdimension	Quality Dimension	Factors to Consider
Source	Institutional/business environment	Sustainability of the entity-data provider Reliability status Transparency and interpretability
	Privacy and security	Legislation Data keeper vs. data provider Restrictions Perception
Metadata	Complexity	Technical constraints Whether structured or unstructured Readability Presence of hierarchies and nesting
	Completeness	Whether the metadata is available, interpretable and complete
	Usability	Resources required to import and analyze Risk analysis
	Time-related factors	Timeliness Periodicity Changes through time

Table V.1. (continued)

Hyperdimension	Quality Dimension	Factors to Consider
Metadata	Coherence-linkability	Presence and quality of linking variables Linking level
	Coherence-consistency	Standardization Metadata available for key variables (classification variables, construct being measured)
	Validity	Transparency of methods and processes Soundness of methods and processes
Data	Accuracy and selectivity	Total survey error approach Reference datasets Selectivity
	Coherence-linkability	Quality of linking variables
	Coherence-consistency	Coherence between metadata description and observed data values
	Validity	Coherence between processes and methods and observed data values

Source: UNECE (2014).

Not shown in the table but presented in the text are a number of possible indicators for each quality dimension. The indicators are mostly posed as questions, but some of the indicators specify calculations. For example, under coherence-linkability in the Data hyperdimension, two indicators are listed:

- Are potential linking variables present on the file that could be used for data integration with other data files?
- Calculate the percentage of units linked and not linked in both the Big Data and other data sources. The indicator is the percentage of units linked unambiguously (strong link) divided by the percentage of units linked with a soft link (linking requirements were relaxed in order to link more units)

While the quality dimension coherence-linkability appears under both the metadata and data hyperdimensions, only this one set of possible indicators is offered. In addition, we are puzzled

that the second indicator presumes that linkage has already occurred. We would not expect such linkage to be included in the input phase.

A more extensive set of possible indicators is presented under accuracy and selectivity:

- If a reference dataset is available, assess coverage error. For example, measures of distance between Big Data population and the target population (for example, Kolmogorov-Smirnov Index, Index of Dissimilarity)
- Does the file contain duplicates?
- Are the data values within the acceptable range?
- Assessment (also qualitative) of sub-populations that are known to be under/over-represented or totally excluded by Big Data source
- Assessment of spatial distribution of measurement instrument and of periodicity of observations
- Selectivity: Derive an R-index for unit composition

The presentation of these as “possible” indicators suggests that the team had only begun to lay out these indicators and that continued development can be expected.

For the throughput phase, which encompasses the span between acquisition of the data and dissemination of a final product, the authors of the quality framework depart from the presentation of a configuration of hyperdimensions and dimensions and possible indicators. Rather, they present some general principles, the most significant of which from a quality perspective is the idea of “quality gates.” A quality gate is a checkpoint at which the quality of data is assessed. Both the measures and the locations of the quality gates are determined in advance. Quality gates are more substantial than quality checks. A given quality gate may involve multiple dimensions of quality, with different sets of dimensions applying to different gates. In summarizing their assessment of throughput quality, the authors observe that “it is not sufficient to simply expand our understanding of data quality to a wider range of data formats and sources.” Instead, “more general conceptions of data quality must be developed that

encapsulate new techniques as well as old, and that are flexible enough to be applicable to the full range of outputs and products that are possible from Big Data.”

For the output phase the quality framework focuses on the information that a consumer of the data would ideally like to have. Table V.2 summarizes the output phase dimensional

Table V.2. Dimensional structure of the output phase of the UNECE Big Data quality framework

Hyper-dimension	Quality Dimension	Factors to Consider
Source	Institutional/business environment	Type of data source Arrangements and quality assurance Type of use of the Big Data source
	Privacy and security	Legislation Actual limitations in the use of data Actions undertaken
Metadata	Complexity	Data treatment; output limitations
	Accessibility and clarity	Data and metadata accessibility Clear definitions, explanations Conformity to standards Presence of hierarchies and nesting
	Relevance	Extent to which the data measures the concepts meant to be measured for its intended uses
Data	Accuracy and selectivity	Traditional measures of accuracy Selectivity
	Validity	Correlation with similar metrics Utility Conceptual soundness
	Coherence-linkability	
	Coherence-consistency	
	Time-related factors	Timeliness Periodicity

Source: UNECE (2014).

structure. Here the source hyperdimension has two quality dimensions; the metadata hyperdimension has three; and the data hyperdimension has five.

The authors note that the output quality dimensions tend to be more holistic than those of input or throughput quality and that, for this reason, specific indicators of output quality tend to be less useful. The dimensions of coherence-linkability and coherence-consistency under the data hyperdimension have no factors to consider and, therefore no quality indicators. The indicators under accuracy and selectivity are a subset of those listed for this dimension in the input phase. These and other features underscore the extent to which the quality framework for Big Data is a work in progress.

B. IMF Internal Group on Big Data

In August 2016 the IMF established an Internal Group on Big Data within its statistics department with the objective of investigating “opportunities and challenges of Big Data for macroeconomic and financial statistics” (Hammer et al. 2017). Big Data can benefit official statistics by:

- Answering new questions and producing new indicators
- Bridging time lags in the availability of official statistics and supporting the timelier forecasting of existing indicators
- Providing an innovative data source in the production of official statistics

At the same time, it should be noted that the opportunities that Big Data afford for macroeconomic and financial statistics vary across statistical domains. The most promising opportunities lie in “flows and transactions, insights, correlations, trends, and sentiments.” Big Data appear to offer less for “statistics on stocks or the breakdown of flows into transactions, revaluations, and other volume changes.” Moreover, the challenge that Big Data present for comparability of economic statistics across countries and over time must be addressed.

An observation of the IMF group that speaks directly to the efforts of the FCSM working group is that “official statistics need to develop new data quality concepts and expand existing frameworks to incorporate the opportunities and challenges that come with Big Data.” In addition, certain obligations attend efforts to exploit data sources as novel as those provided by Big Data. In particular, “the use of Big Data for new indicators must be made transparent in terms of the applied methodology and the data origin; otherwise the value of policy advice and forecasting can be seriously weakened.” This plea for transparency in the use of Big Data in official statistics underscores a major focus of the FCSM and OMB in addressing the implications of integrated data.

Different uses of Big Data may demand different approaches to quality assessment. On this point Hammer (2017) contrasts the use of Big Data to uncover insights, trends, and sentiments with the use of Big Data in official statistics. However, both types of uses will require consistent and harmonized historical time series.

C. AAPOR Big Data Task Force

As a professional association “dedicated to advancing the study of ‘public opinion,’” AAPOR’s goals include working to improve data collection, helping to make its members and various constituencies better users of surveys and survey findings, and keeping them informed about new developments in the field (AAPOR Big Data Task Force 2015). Against this backdrop AAPOR’s council saw a need to address a number of issues related to Big Data and convened a task force to prepare a report that would describe both the potential of Big Data and the challenges confronting its use, present potential solutions, and identify key research needs. In addition to representatives of both producers and users of survey data in the U.S., the Task Force included members from two European universities and a national statistical organization. The report includes a relevant discussion of data quality, summarized below.

The Task Force noted that for survey data both sampling and non-sampling error have been expressed in a very useful fashion in the TSE framework, which we discussed in Chapter III. The Task Force concluded that a total error framework is needed for Big Data, and it offered “a skeletal view” of such a framework. The TSE framework is sufficiently general that it can be applied to any dataset that conforms to the row/column format of survey data, where rows represent elements of a sample or population, columns represent characteristics of the row elements, and cells hold the values of these characteristics for each row element. Total error is the sum of errors at the row, column, and cell level. Row error derives from deficiencies in the representation of the target population; column error derives from deficiencies in measuring the characteristics of the row elements due, for example, to mislabeling or bias; and cell error derives from incorrect or missing measurement of the column characteristics. When Big Data has a row/column structure, or such a structure can be imposed on the data, total error can be evaluated in this same manner. Where Big Data will differ from survey data is in the composition of the error. At the row level, sampling error may be minimal or nonexistent, but undercoverage and overcoverage may abound. At the column level, error may be dominated by deviations of measured characteristics from what the analyst wishes to observe. At the cell level, rates of missing data and inaccurate measurement may be high. The Task Force concludes, though, that to date, “very little effort has been devoted to enumerating the error sources and the error generating processes for Big Data.”

VI. DISCUSSION AND CONCLUSIONS

The goal of this review was to compile information on international standards and guidelines on quality reporting relevant to statistical estimates that combine multiple sources of data. The information presented in this report is intended to serve as a resource to the FCSM Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources. In this concluding chapter we present highlights from the review that we believe will address the working group's needs most directly.

The European context

Standard and guidelines issued by Eurostat and the ESS reflect the need for comparability in the statistics produced by the member states of the European Union for their respective populations. Manuals defining appropriate methods for the production of economic and demographic statistics for European nations are a fundamental part of the environment in which the national statistical agencies operate. The quality standards and guidelines from the European Union make frequent reference to European statistical standards and methods. This is not to say that the individual nations of the European Union do not have their own quality frameworks and guidelines; we have noted some of the unique features of the standards and guidelines implemented by selected national statistical agencies. But there is considerable similarity across countries that derives from their joint membership in the European Union. There is no parallel to this in the U.S. outside of some of the major macro-economic statistics like Gross Domestic Product, where international comparisons are common.

The quality concept

The concept of quality as expressed in a wide array of quality frameworks for statistical data is characterized by several features:

- Quality is commonly defined as fitness for use.

- Quality is multi-dimensional; five dimensions appear almost universally in quality frameworks around the world: (1) relevance, (2) accuracy and reliability, (3) timeliness and punctuality, (4) coherence and comparability, and (5) accessibility and clarity.
- These dimensions can be mutually reinforcing. For example, accuracy, timeliness, and accessibility can enhance the relevance of a statistic while declines in accuracy, timeliness, or accessibility can make a statistic less relevant.
- There are also trade-offs among the dimensions. For example, improved timeliness may require some sacrifice of accuracy or, conversely, improved accuracy may necessitate a reduction in timeliness.
- Indicators of accuracy and reliability tend to be quantitative while indicators of the other dimensions tend to be qualitative.
- A number of other dimensions appear in some national statistical organizations' quality frameworks; examples include interpretability, credibility, methodological soundness, serviceability, assurances of integrity, and confidentiality.
- Granularity, promoted as a dimension of quality by the recent CNSTAT panel on multiple data sources, is cited only rarely in international quality frameworks and supporting documents; Statistics Netherlands includes subpopulation detail as one of several additional dimensions of quality in its checklist for statistical output, and the ABS lists geographic detail as a factor in assessing relevance.

While the multi-dimensional formulation of quality suggests comparable importance among the dimensions, discussions of quality in the international literature give disproportionate attention to accuracy. Notably, nearly half of the main text of the *ESS Handbook for Quality Reports* is devoted to accuracy and reliability. And while statistical uses of administrative records date back centuries in Europe, more than half of the Handbook's discussion of this dimension is focused on sample surveys.

Standards for integrated data

With respect to standards for integrated data we find that:

- Only one national statistical organization—Stats NZ—has developed a quality framework explicitly designed to address integrated data
- Eurostat's quality standards and guidelines, which apply to most of Europe and are perhaps the most extensive, deal with integrated data to a much more limited degree
- Efforts to deal with quality aspects of administrative data are much farther along than efforts to deal with the quality of other forms of Big Data

Stats NZ (2016), as noted, has addressed integrated data most directly, building on Zhang's (2012) adaptation of the TSE model. Where the TSE model follows the life cycle of a survey and culminates in a single survey statistic, Zhang proposed a "life-cycle model of integrated statistical micro data" that culminates in an entire dataset. Zhang's model has two phases. The first phase describes a single data source, but each input to the integrated micro data, whether a survey or administrative data source, has its own phase one assessment. For phase one, each data source is assessed relative to its original purpose. Phase two describes the integration of these multiple sources to create a new micro data source. The sources of error depicted in phase two reflect the integration process, which may include transformation of the input data to match the concepts (measures and population) that define the integrated data.

Zhang's two-phase framework for integrated data is incorporated in Stats NZ's *Guide to Reporting on Administrative Data Quality* (Stats NZ 2016). The Guide includes quality indicators for each of the phase one and phase two error sources depicted in Zhang's framework. There are both quantitative and qualitative indicators for phase one but only quantitative indicators for phase two. The 19 quantitative quality indicators for phase two address coverage error, record linkage methods and results, and other sources of error in representation of the target population and measurement of target concepts. All of the indicators but especially those for phase two would merit close review by the FCSM working group.

Reid et al. (2017) add a third phase to Zhang's framework in order to provide for assessing the quality of the statistical estimates that are derived from the integrated micro data. Phase three returns the focus of the framework to the single estimate that is the endpoint of the TSE model. Quality indicators for phase three have not been defined as yet—in part because the statistical methods used to generate the final estimates are varied. To underscore this point, Reid et al. present three case studies that demonstrate different approaches to evaluating estimates produced

from integrated data. The first case study involved the redesign of a survey to incorporate modeled values from administrative data. The second case study involved the prospective replacement of personal income measures in a household survey with data from linked tax records. The third study involved design of an approach to population estimation based on linked administrative sources. Case study two illustrates a problem likely to occur in the substitution of administrative data for survey data: the administrative variable is biased whereas the survey estimate may be unbiased but has substantial measurement error. Case study three highlights issues that arise when combining a set of overlapping administrative datasets that individually capture only part of the total population.

The distinction between the original purpose of an administrative data source and its statistical use as one of multiple sources in an integrated dataset is discussed repeatedly, albeit in different ways. For example, the ESS Handbook contrasts the concepts or definitions embedded in the data, which are fixed, and those desired by users, which may vary with each new use. The OECD (2011) specifies that the documentation for statistics derived from administrative sources should include the purposes for which the administrative data were originally collected and the merits and shortcomings of these data relative to the statistical purpose to which they have been applied. In Zhang's two-phrase framework for integrated data, the dimension of relevance is given a new meaning, referring to the appropriateness of measures obtained from administrative data when used as alternatives or supplements to survey-based measures.

Continuing on this issue, various quality frameworks specify a detailed review of an administrative data source before its use for statistical purposes. Statistics Canada, a world leader in the substitution of administrative records for survey responses, advises that the decision to use administrative records in conjunction with a survey be preceded by a detailed assessment of such

records that addresses the quality dimensions of relevance, accuracy, timeliness, and coherence (Statistics Canada 2009). If the data are used, the results of such a review should be reported.

If and when such reviews occur in the U.S., their results are rarely reported. More consistent reporting of the results of these reviews would be consistent with greater transparency.

Publication of the results of commonly used administrative data sources would also reduce duplication in the performance of these reviews and contribute to a better informed community of users.

With regard to the quality of the statistical outputs generated from multiple data sources, the Handbook's principal recommendation applies only when both preliminary and revised estimates are produced, in which case the magnitudes of the revisions can be informative about quality.

The Handbook notes as well that measures of statistical precision such as coefficients of variation should reflect the composite estimation. Statistics Canada, with its extensive experience in combining survey responses and administrative data, has developed procedures for these types of calculations. The Handbook includes a discussion of non-probability sampling and the need to account for it in estimates of precision while acknowledging that there is no generally agreed-upon approach.

Issues in quality measurement for integrated data

Combining multiple data sources creates a number of issues for quality measurement, which arise from the application of particular statistical methods. These issues are discussed in a number of the sources we reviewed, with the most attention afforded by the ESS Handbook.

Quality measurement of integrated data will necessitate the development of measures that focus on non-sampling error. With administrative data and Big Data, the importance of sampling error is greatly diminished while the importance of non-sampling error is elevated. The Handbook's discussion of non-sampling error includes coverage error, measurement error,

nonresponse error, and processing error. Coverage error encompasses undercoverage, overcoverage, and duplication. A performance indicator reflecting overcoverage is recommended for inclusion in quality reports, but no indicator is recommended for undercoverage, which is acknowledged as the most challenging to measure. No performance indicators are provided for measurement error or processing error, and only the aforementioned response rates are suggested for nonresponse, although some additional descriptive information on response patterns and a qualitative treatment of the risk of bias should be included in quality reports. Qualitative assessments of measurement error and processing error are recommended as well.

The integration of multiple sources is likely to require record linkage. The quality of the combined data will depend in an important way on the quality of the linkage. Consequently, indicators of the quality of the linkage may become as important to integrated data as response rates are to survey data. First, an assessment of the quality of the unit identifiers in each data source should be included in any assessment of these data sources prior to their use. Second, measures of the quality of the record linkage between each pair of sources should be generated as part of documenting the impact of combining data sources on the quality of the resulting estimates. The false negative match rate (failure to link two records that refer to the same entity) is conceptually analogous to the survey nonresponse rate, but the nonresponse rate can be measured directly with data collected in conducting the survey. Unless there is independent information as to which unmatched records should have matched, the false match rate cannot be measured except indirectly, through an evaluation of the linkage methodology applied to a dataset for which the expected match rate absent any errors is known. The false positive rate has no counterpart in response rates, either in form or its implications for quality, but it can be estimated through a review of observed matches.

Modeling also has a critical role to play in the development of integrated data. Modeling is addressed most extensively in the ESS Handbook's recommendations on quality reporting. This discussion occurs under the topic of general issues in the discussion of the dimension of accuracy and reliability. When modeling plays a role in estimation, the model, its assumptions, and its validity for that specific application should be discussed in the quality report. No indicators are proposed, perhaps because modeling can assume many forms, but extensive descriptive information is requested. The need to describe modeling assumptions is echoed in other quality frameworks and standards for reporting.

Imputation may be considered a type of modeling. Regardless of how imputation is characterized, however, its importance has grown with increasing item nonresponse. While U.S. surveys include indicators of imputation in their public use files, imputation rates are rarely reported. The OMB standards and guidelines for surveys mandate the inclusion of such indicators, but they do not request that rates of imputation be reported (OMB 2006). International standards commonly do specify the reporting of imputation rates, and the ESS Handbook extends this to include a discussion of the methods of imputation and what is known about their effects on the estimates. Greater use of imputation may not be a uniform property of integrated data, but more frequent reporting of imputation rates—especially for statistical estimates with high rates of imputation—would increase transparency in the reporting of data quality.

The impact of methods of statistical disclosure control on the quality of statistical estimates is addressed in the UK *Guidelines for Measuring Statistical Output Quality* (Office for National Statistics 2013), which includes several measures of the impact of statistical disclosure control on accuracy and reliability as well as relevance, coherence and comparability, and accessibility and clarity. This topic received little attention elsewhere but can be expected to grow in

importance with greater use of administrative data and possibly other forms of Big Data and with more frequent production of estimates combining multiple data sources.

Finally, an issue that arises in the quality literature but is not explicitly addressed is the extent to which quality can be measured usefully at the dataset level—as in Zhang’s (2012) two-phase framework—or should be restricted to the individual estimate—as in the TSE model of Groves et al. (2009). As we noted, Reid et al. (2017) of Stats NZ added a third phase to Zhang’s framework to enable quality measurement of integrated data at the level of the individual estimate. Certainly, there are aspects of data integration—such as record linkage—for which the most appropriate measures of quality apply to the entire dataset. But, in the end, quality is rarely uniform across the variables of a dataset, and some if not many of the benefits of combining multiple data sources—such as reduction in measurement error or improved imputation—affect individual variables more than the dataset as a whole.

Quality and Big Data

The quality assurance frameworks were designed primarily for use with survey data, with more limited attention to administrative data—and generally in the form of registers. The frameworks will require considerable adaptation to be applied productively to many forms of Big Data, whether such data are being used alone or, more likely, in combination with other sources. Quality frameworks for survey data reflect the statistical agency’s control over every aspect of the survey design, data collection, processing, analysis, and dissemination. With this control comes a detailed understanding of how the data were created, which can be expressed in correspondingly detailed metadata. This is less true of administrative data, which is likely to have documentation adequate for administrative use by a dedicated community of administrative agency users but generally not sufficient for research use. By contrast, organic or found data may have been generated with little or no control beyond the placement of a collecting or measuring

tool. There may be limited or no metadata available about even the most structured forms of Big Data. Documentation on privately collected Big Data, such as may exist, is likely to be proprietary. Moreover, the size of Big Data files may demand sophisticated computing technology and relevant institutional knowledge to analyze, compounding the difficulty of calculating many of the types of quality indicators reviewed in this report.

Efforts to develop quality assurance frameworks and data quality standards for Big Data are recent and in the early stages of development. The most significant effort in this area, *A Suggested Framework for the Quality of Big Data*, was produced by the UNECE Big Data Quality Task Team (UNECE 2014). After studying existing quality frameworks designed for survey and administrative data, the team concluded that such frameworks would be inadequate for Big Data because of the extensive scope of the latter. The suggested framework includes 11 quality dimensions, which are also present in one or more traditional quality frameworks. Possible indicators of each dimension are mostly posed as questions rather than quantitative measures, although there are a few exceptions such as linkage rates, coverage measures and an R-index to measure representativeness. The quality framework is clearly a work in progress and will continue to evolve.

Quality reporting

Statistics Canada's *Policy on Informing users of Data Quality and Methodology*, which dates back to 2000, is a strong expression of the value of transparency, which the FCSM working group may want to review.

Many of the quality assurance frameworks and the accompanying standards and guidelines reviewed in this report are associated with extensive prescriptions for quality assessments and their communication to data users in detailed quality reports. Notably, the volume and types of information requested in Eurostat quality reports bear substantial resemblance to what was

included in the quality profiles prepared by a number of U.S. federal agencies in the 1990s and early 2000s. A survey quality profile summarizes what is known about the sources and magnitudes of errors in a survey; it provides a systematic and comprehensive review across the spectrum of survey activities in which both qualitative and quantitative results are brought together to allow an assessment of the quality of the survey operations and the data (Kasprzyk and Kalton 2001).²³ While quality profiles were intended for recurring surveys, they were updated or repeated for only one survey—the Census Bureau’s Survey of Income and Program Participation—and no new profile has been produced in the past decade. There are a number of reasons why the preparation of quality profiles has not continued. Their production demands resources that are increasingly less available, they require detailed information that may not exist, and their value to the survey producer in terms of suggesting future improvements is questionable. This prior experience suggests that federal agencies are not likely to embrace the recommendations of international agencies for substantially more extensive reporting on quality than is done currently. We suspect that a more popular format may be one similar to the Source and Accuracy statements that appear as appendices in some Census Bureau publications.

²³ Quality profiles tended to focus on the accuracy dimension of quality.

REFERENCES

- AAPOR Big Data Task Force. "Big Data in Survey Research." *Public Opinion Quarterly*, vol. 79, no. 4 (Winter 2015), pp. 839-880.
- Australian Bureau of Statistics. *The Australian Bureau of Statistics Data Quality Framework*. Canberra, 2009.
- Bakker, Bart F.M. "Micro-integration: State of the Art." In ESSnet on Data Integration. Draft Report of WP1: State of the Art on Statistical Methodologies for Data Integration, 2010, pp. 58-81.
- Bank of England. Data Quality Framework. Bank of England, Statistics and Regulatory Data Division, 2014.
- Biemer, Paul, and Lars Lyberg. *Introduction to Survey Quality*. New York: John Wiley & Sons, 2003.
- Brackstone, Gordon. "Managing data quality in a statistical agency." *Survey Methodology*, vol. 25, no. 2 (December 1999), pp. 139-149.
- Brick, J. Michael and Douglas Williams. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *Annals of the American Academy of Political and Social Science*, volume 645 (January 2013), pp. 36-59.
- Citro, Connie. "From Multiple Modes for Surveys to Multiple Data Sources for Estimates." *Survey Methodology*, vol. 40 (no. 2, 2014), pp. 137-161.
- Daas, Piet, et al. *Deliverable 4.1: List of Quality Groups and Indicators Identified for Administrative Data Sources*, Report for Work Package 4 of the European Commission 7th Framework Program BLUE-ETS. Brussels, Belgium: European Commission, 2011.
- Daas, Piet, and Saskia Ossen. *Deliverable 4.2: Report on Methods Preferred for the Quality Indicators of Administrative Data Sources*, Report for Work Package 4 of the European Commission 7th Framework Program BLUE-ETS. Brussels, Belgium: European Commission, 2011.
- Daas, Piet, Saskia Ossen, Rachel Vis-Visschers, and Judit Arends-Toth. "Checklist for the Quality Evaluation of Administrative Data Sources." Discussion paper 09042. The Hague: Statistics Netherlands, 2009. <http://ec.europa.eu/eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683-4477be82ee60>
- Daas, Piet J. H., Marco J. Puts, Bart Buelens, and Paul A. M. van den Hurk. "Big Data as a Source for Official Statistics." *Journal of Official Statistics*, vol. 31, no. 2 (June 2015), pp. 249-262.

- De Leuw, Edith, and Wim de Heer. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In Robert M. Groves, Don A. Dillman, John L. Eltinge, and Rod J. A. Little (Eds.), *Survey Nonresponse*. New York: Wiley, 2002.
- De Smedt, Marleen. "Invited Commentary, Special Section: Addressing the Needs of Official Statistics Users: The Case of Eurostat." *Journal of Official Statistics*, vol. 32, no. 4 (December 2016), pp. 913-916.
- European Commission. *European Statistics Code of Practice*. Brussels, Belgium: European Statistical System Committee, September 28, 2011.
- European Statistical System Committee. *Quality Assurance Framework of the European Statistical System, Version 1.2.*, 2015.
<http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646>
- Eurostat. *ESS Handbook for Quality Reports*, 2014 edition. Luxembourg: Publications Office of the European Union, 2015. <http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf/18dd4bf0-8de6-4f3f-9adb-fab92db1a568>
- Eurostat. *Quality report of the European Union Labour Force Survey, 2015*. Luxembourg: Publications Office of the European Union, 2017.
- Federal Committee on Statistical Methodology. *Measuring and Reporting Sources of Error in Surveys*. Statistical Policy Working Paper 31. Office of Management and Budget, 2001.
<https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/04/spwp31.pdf>
- Fellegi, Ivan P., and Alan B. Sunter. "A Theory for Record Linkage." *Journal of the American Statistical Association*, vol. 64 (1969), pp. 1183-1210.
- Groves, Robert M., Floyd J. Fowler, Jr., Mick Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey Methodology. Revised edition*. New York: Wiley, 2009.
- Hammer, Cornelia L. "IMF Strategy to Multi-source Information with Big Data." Presentation to the 4th UN Conference on Big Data, Bogota, Colombia, November 8 to 10, 2017.
- Hammer, Cornelia L., Diane C. Kostroch, Gabriel Quiros, and STA Internal Group. "Big Data: Potential, Challenges, and Statistical Implications." *IMF Staff Discussion Note 17/06*. International Monetary Fund, September 2017.
- Holmberg, Anders, and Christine Bycroft. "Statistics New Zealand's Approach to Making Use of Alternative Data Sources in a New Era of Integrated Data." In Paul P. Biemer et al., editors, *Total Survey Error in Practice*. New York: John Wiley and Sons, 2017.
- Horrigan, Michael W. "Big Data: A BLS Perspective." *Amstat News*, no. 427 (January 2013), pp. 25-27.

International Monetary Fund. *Data Quality Assessment Framework and Data Quality Program*. IMF, 2003. <http://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm>

Iwig, William, Michael Berning, Paul Marck, and Mark Prell. *Data Quality Assessment Tool for Administrative Data*. Washington, DC: Federal Committee on Statistical Methodology, February 2013.

Juran, Joseph M., and Frank M. Gryna, Jr. *Quality Planning and Analysis: From Product Development through Use*. New York: McGraw-Hill, 1980.

Kasprzyk, Daniel, and Graham Kalton. "Quality Profiles in U.S. Statistical Agencies." *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm, Sweden, 2001.

Kilss, Beth, and Frederick J. Scheuren. "The 1973 CPS-IRS-SSA Exact Match Study." *Social Security Bulletin*, vol. 51, no. 7 (October 1978), pp. 14-22.

Laitila, Thomas, Anders Wallgren, and Britt Wallgren. "Quality Assessment of Administrative Data." *Research and Development—Methodology Reports from Statistics Sweden 2011:2*. Stockholm, Statistics Sweden, 2011.

Laney, Douglas. "The Importance of 'Big Data': A Definition." Gartner Inc., 2012.

Lavallee, Pierre. "Combining Survey and Administrative Data: Discussion Paper." *Proceedings of the Second International Conference on Establishment Surveys*. Alexandria, VA: American Statistical Association, 2000.

Lavallee, Pierre. "Quality Indicators when Combining Survey Data and Administrative Data." *Proceedings of the XXII International Methodology Symposium*. Ottawa: Statistics Canada, 2005.

Lohr, Sharon L., and Trivellore E. Raghunathan. "Combining Survey Data with Other Data Sources." *Statistical Science*, vol. 32, no. 2 (2017), pp. 293-312.

National Academies of Sciences, Engineering, and Medicine. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods: Frameworks, Methods, and Assessment, Robert M. Groves and Brian Harris-Kojetin (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2017a. <https://doi.org/10.17226/24652>

- National Academies of Sciences, Engineering, and Medicine. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods: Frameworks, Methods, and Assessment, Robert M. Groves and Brian Harris-Kojetin (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2017b.
<https://doi.org/10.17226/24893>
- National Research Council. *Nonresponse in Social Science Surveys: A Research Agenda*. Panel on a Research Agenda for the Future of Social Science Data Collection. Roger Tourangeau and Thomas J. Plewes (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2013.
- Nelson, Nils, and Kirsten West. "Interview with Lars Thygesen." *Statistical Journal of the IAOS*, vol. 30, no. 2 (2014), pp. 67-73.
- Office for National Statistics. *Guidelines for Measuring Statistical Output Quality*. Version 4.1. United Kingdom, September 2013.
- Organization for Economic Cooperation and Development. *Quality Framework and Guidelines for OECD Statistical Activities*. Version 2011/1. Paris: OECD, 2012.
<http://www.oecd.org/dataoecd/26/38/21687665.pdf>
- Reid, Giles, Felipa Zabala, and Anders Holmberg. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." *Journal of Official Statistics*, vol. 33, no. 2 (2017), pp. 477-511.
- Schouten, Barry, Fannie Cobben, and Jelke Bethlehem. "Indicators of Representativeness of Survey Nonresponse." *Survey Methodology*, vol. 35, no. 1 (June 2009), pp. 101-113.
- Statistics Canada. "Policy on Informing Users of Data Quality and Methodology." Ottawa: Statistics Canada, 2002.
- Statistics Canada. *Statistics Canada Quality Guidelines*. Fifth Edition. Publication 12-539-X. Ottawa: Statistics Canada, 2009.
- Statistics Canada. *Statistics Canada's Quality Assurance Framework*. Third edition. Ottawa: Statistics Canada, 2017.
- Statistics Finland. *Guidelines on Professional Ethics*. Revised Edition. Handbooks 30b. Helsinki: Statistics Finland, 2006.
- Statistics Finland. *Quality Guidelines for Official Statistics*. 2nd Revised Edition. Helsinki: Statistics Finland, 2007.
- Statistics Netherlands. *Quality Guidelines 2014: Statistics Netherlands' Quality Assurance Framework at Process Level*. The Hague: Statistics Netherlands, 2014.
<https://www.cbs.nl/en-gb/background/2014/12/quality-guidelines-2014>

- Statistics New Zealand. *Guide to Reporting on Administrative Data Quality*. Wellington, NZ: Statistics New Zealand, 2016.
- Statistics Sweden. *Official Statistics of Sweden—Annual Report 2016*. Stockholm: Statistics Sweden, 2017.
- Tam, Siu-Ming, and Frederic Clarke. “Big data, official statistics, and some initiatives by the Australian Bureau of Statistics.” *International Statistical Review*, vol. 83, no. 3 (December 2015), pp. 436-448.
- Trepanier, Julie, Claude Julien, and John Kovar. “Reporting Response Rates when Survey and Administrative Data are Combined.” *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. Arlington, VA, November 14-16, 2005.
- UK Statistics Authority. *Exposure Draft of a Report from the UK Statistics Authority: Quality Assurance and Audit Arrangements for Administrative Data*. July 2014.
- UK Statistics Authority. *Quality Assurance of Administrative Data: Setting the Standard*. January 2015a.
- UK Statistics Authority. *Administrative Data Quality Assurance Toolkit*. January 2015b.
- UK Statistics Authority. *Code of Practice for Statistics: Ensuring Official Statistics Serve the Public*. Edition 2.0. February 2018.
- United Nations Economic and Social Council, Statistical Commission. “Report of the Global Working Group on Big Data for Official Statistics.” United Nations, December 2015.
- United Nations Economic Commission for Europe. *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. Geneva, Switzerland: United Nations Publication, 2011.
- United Nations Economic Commission for Europe. *Classification of Types of Big Data*. UNECE, 2013.
<http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data>
- United Nations Economic Commission for Europe. *A Suggested Framework for the Quality of Big Data*. UNECE, December 2014.
<http://www1.unece.org/stat/platform/display/bigdata/2014+Project>
- U.S. Office of Management and Budget. *Statistical Policy Handbook*. Washington, DC: Office of Management and Budget, 1978.
- U.S. Office of Management and Budget. *Statistical Policy Directive No. 2: Standards and Guidelines for Statistical Surveys*. Washington, DC: Office of Management and Budget, 2006.
- Van Norderpelt, Peter. *Checklist Quality of Statistical Output*. The Hague: Statistics Netherlands, 2009. <https://www.cbs.nl/en-gb/background/2009/07/checklist-quality-of-statistical-output>

Van Nederpelt, Peter. "A New Model for Quality Management." Discussion paper 201017. The Hague: Statistics Netherlands, 2010.

Wallgren, Anders, and Britt Wallgren. *Register-based Statistics: Administrative Data for Statistical Purposes*. Hoboken, NJ: John Wiley & Sons, Inc., 2007.

Zhang, Li-Chun. "A Unit-Error Theory for Register-Based Household Statistics." *Journal of Official Statistics*, vol. 27, no. 3 (September 2011), pp. 415-432.

Zhang, Li-Chun. "Topics of statistical theory for register-based statistics and data integration." *Statistica Neerlandica*, vol. 66, no. 1 (2012), pp. 41-63.

www.mathematica-mpr.com

***Improving public well-being by conducting high quality,
objective research and data collection***

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■
SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD



Mathematica® is a registered trademark
of Mathematica Policy Research, Inc.