



A Framework for Data Quality

FCSM-20-04

September 2020

2020 Federal Committee on Statistical Methodology Members

Stephen Blumberg

National Center for Health Statistics

Chris Chapman

National Center for Education Statistics

Jennifer Edgar

Bureau of Labor Statistics

John Eltinge

U.S. Census Bureau

John Finamore

National Science Foundation

Dennis Fixler

Bureau of Economic Analysis

Michael Hawes

U.S. Census Bureau

Jennifer H. Madans

National Center for Health Statistics

Rochelle (Shelly) Martinez (Chair)

Office of Management and Budget

Wendy Martinez

Bureau of Labor Statistics

Jaki McCarthy

National Agricultural Statistic Services

Jennifer Nielsen (Secretary)

National Center for Education Statistics

Jennifer M. Ortman

U.S. Census Bureau

Anne Parker

Internal Revenue Service

Jennifer D. Parker

National Center for Health Statistics

Polly Phipps

Bureau of Labor Statistics

Mark Prell

Economic Research Service

Joseph Schafer

U.S. Census Bureau

Rolf R. Schmitt

Bureau of Transportation Statistics

Marilyn M. Seastrom

National Center for Education Statistics

Joy Sharp

Energy Information Administration

Robert Sivinski

Office of Management and Budget

G. David Williamson

Agency for Toxic Substances and Disease Registry

Linda J. Young

National Agricultural Statistic Service

Workgroup on Transparent Reporting of Data Quality

(In alphabetical order)

Keenan Dworak-Fisher

Bureau of Labor Statistics

Lisa Mirel

National Center for Health Statistics

Jennifer D. Parker

National Center for Health Statistics

John Popham

Bureau of Justice Statistics

Mark Prell

Economic Research Service

Rolf R. Schmitt

Bureau of Transportation Statistics

Marilyn M. Seastrom

National Center for Education Statistics

Linda J. Young

National Agricultural Statistic Service

Recommended citation: Federal Committee on Statistical Methodology. 2020. *A Framework for Data Quality*. FCSM 20-04. Federal Committee on Statistical Methodology. September 2020.

A Framework for Data Quality

Contents

- Executive Summary 1
- Preface 9
- Acknowledgements 10
- List of Acronyms..... 11

- 1. Introduction 12
 - Index..... 15
 - Legend..... 15
- 2. A Unified Data Quality Framework 17
- 3. Factors that Affect Data Quality..... 30
- 4. Best Practice for Identifying and Reporting Data Quality..... 48
- Appendix A. Additional Background on Data Quality 57
- Appendix B. Accuracy and Reliability of Integrated Data..... 67

Executive Summary

Effective understanding of data quality is essential for public officials, private businesses, and the public to make data-driven decisions. New sources of data, uses of existing data, analysis methods, and increasing reliance on integrating data from multiple sources bring new opportunities and challenges to federal agencies that provide information to support public and private decisions. Although new data sources and methods show great promise, the quality of these data must be evaluated. Inferior quality data can result in misleading information and poor decisions.

A clear, documented understanding of the strengths and weaknesses of data assures attention to ameliorating the weaknesses, enables appropriate uses of the data, and reinforces the credibility of the data and their use. All data have strengths and weaknesses across the multiple dimensions of data quality. These strengths and weaknesses typically involve trade-offs of comprehensive coverage and accuracy versus timeliness and other quality dimensions. Data users who understand the fitness-for-use of data are more likely to use them appropriately, whether for secondary use in developing other data products, for conducting data analysis, or when using data outputs for decision making.

The Interagency Council on Statistical Policy (ICSP) has indicated that “agencies should work to adopt a common language and framework for reporting on the quality of data sets and derivative information they disseminate” (ICSP 2018). This report presents a framework for identifying data quality for all data, summarizes the current state of practice in identifying threats to data quality for the components of the framework, and provides guidance for promoting effective reporting of data quality. Statistical agencies in many countries have extensive, well-established methods for identifying and reporting threats to quality in data collected and designed for statistical purposes, particularly sample surveys. Methods are less well-developed for dealing with threats to quality from sources other than surveys, such as administrative records and readings from sensors, and other data originally collected for nonstatistical purposes.

Background and Purpose

In response to the rapidly changing world of data sources and analysis methods, the Federal Committee on Statistical Methodology (FCSM) established a Data Quality Analysis Working Group to provide practical information on identifying and reporting data quality for federal agencies. The group’s initial focus was to establish a comprehensive data quality framework that provides an inventory of the elements (i.e., domains and dimensions) of data quality with a review of identifiable threats to each dimension of data quality. Establishing such a framework was one of the key actions recommended by the Committee on National Statistics (CNSTAT) Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation (CNSTAT 2017, 2018). In particular, CNSTAT noted that a useful framework needs to account for different dimensions of quality, such as timeliness and granularity, so that it can be used to consider tradeoffs among multiple dimensions. Other countries’ statistical leaders have defined frameworks to support guidance on identifying and reporting data quality (see e.g., Czajka and Stange 2018). After extensive considerations of international research and standards, the working group established a quality framework based on the 2000 Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000).

The Information Quality Act directed the Office of Management and Budget (OMB) to issue government-wide guidelines that “provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies.” In OMB guidance, information quality is described as an overarching concept that includes Objectivity, Utility, and Integrity of information. This guidance, along with elements that emerged in a commissioned report by Mathematica Policy Research that examined data quality frameworks and standards used outside the United States by national statistical offices and international organizations, including the European Statistical System and a selection of individual European countries, Canada, Australia,

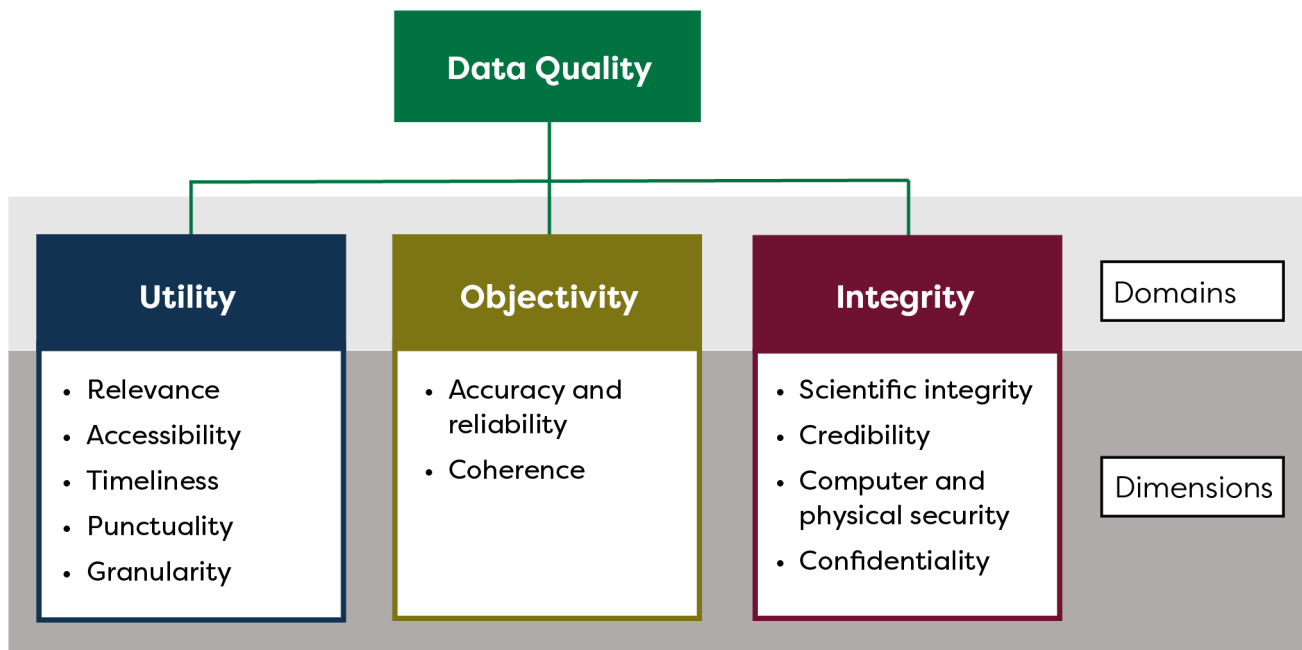
the Organisation for Economic Co-operation and Development (OECD), and the International Monetary Fund (IMF), formed the basis for the FCSM Data Quality Framework.

The FCSM Data Quality Framework provides a common foundation upon which federal agencies can make decisions about the management of data products throughout their lifecycle by identifying and mitigating key data quality threats, evaluating trade-offs among different quality dimensions where necessary, applying accepted methods at an appropriate level of rigor, and accounting for and reporting on the quality of data products and outputs. These activities all support appropriate and effective use of data. Wide application of these practices by all federal agencies is consistent with several OMB policy documents. For instance, OMB memorandum M-19-15, “Guidance on Improving Implementation of the Information Quality Act,” calls on agencies to document and disseminate information on the quality of administrative data that have the potential to be used for statistical purposes, to allow data users to determine the fitness-for-purpose of the data for use in secondary analysis (OMB 2019).

Defining Data Quality

Data quality is the degree to which data capture the desired information using appropriate methodology in a manner that sustains public trust. This definition of data quality, informed by the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) and other sources, was developed by the FCSM for the purpose of the framework. It applies to all data: data collections and data systems; restricted and public use micro-data files; data products produced through data integration, modeling, harmonization and other statistical analyses; and analysis outputs, such as tables, estimates, graphics and reports. Data quality applies to the elements, or components, of data files (e.g., variables, data fields), as well as to the entire data file. It applies to existing methods such as traditional survey design, and new applications of established and emerging methods used to create

Figure ES 1. The FCSM Data Quality Framework



data files, such as artificial intelligence (AI) and machine learning. The definition applies to data produced from different types of input sources, including data collected for nonstatistical purposes, such as data from administrative records and sensors. The definition applies to integrated data products (i.e., record-linked data, modeled data), where integrated data can consist of statistical data, nonstatistical data, or a combination of data sources. Although the term integrated data is used in this report, this type of data are also described by other terms including blended data, multiple source, combined data, and linked data (when applicable).

Components of Data Quality

Data quality in this framework is considered using three broad components, or domains: utility, objectivity, and integrity. Utility refers to the extent to which information is well-targeted to identified and anticipated needs; it reflects the usefulness of the information to the intended users. Objectivity refers to whether information is accurate, reliable, and unbiased, and is presented in an accurate, clear and interpretable, and unbiased manner. Integrity refers to the maintenance of rigorous scientific standards and the protection of information from manipulation or influence as well as unauthorized access or revision.

The framework builds on these three domains, nesting 11 data quality dimensions within the domains, as shown below (Figure ES 1, Table ES 1). The dimensions represent areas in which specific aspects of data quality can be considered. Figure ES 1 illustrates the conception of data quality is quite broad, encapsulating most of the ideals that federal data providers try to uphold. The quality dimensions are defined in Table ES 1. The quality domains and dimensions were informed from multiple published sources and agency experience and defined for use in this framework. This framework differs from data quality frameworks in other disciplines, such as computer and information science, because it focuses on the evaluation of the quality of data in the context of statistical use and decision making, rather than operational purposes.

In other structures used to describe data quality, quality has sometimes been equated primarily with data accuracy. For example, previous FCSM work has focused on various sources of sample survey error and the methods employed to measure and report them (e.g., FCSM 2001). These traditional sources of error remain important to track as the terrain is expanded to include secondary use of nonstatistical data and integrated data. However, accuracy is one of 11 dimensions of data quality in this framework, which highlights the importance of identifying the other dimensions.



Threats to Data Quality

Threats to data quality can be identified for all of the dimensions within the framework, which is an essential step to facilitating their mitigation, managing trade-offs among them, and for reporting data quality. Within the domain of utility, threats to data quality include competing data sources, costs of access and documentation, use of disclosure protections, and delays in data acquisition and processing. Most threats in the domain of objectivity are threats to accuracy and reliability, many of which (e.g., coverage error and nonresponse) are well-documented in the Total Survey Error paradigm (Biemer *et al.* 2017). Although developed for surveys, these threats can be readily applied or adapted to most nonstatistical data. Increasing in importance are threats to accuracy and reliability for integrated data products, such as linkage error, harmonization error, and modeling error. Threats in the domain of integrity include lack of scientific integrity, political interference, and data security failures. Many threats to data quality are relevant for multiple quality domains and dimensions. For instance, use of appropriate statistical methods reduces threats to all domains through dimensions of credibility, accuracy and reliability, and scientific integrity. In some cases, mitigating threats in one area can exacerbate threats in others. Increasing granularity can increase disclosure risks. Timeliness and punctuality, for instance, can be reduced by steps taken to increase accuracy and reliability or by efforts to increase accessibility through documentation and dissemination.

Table ES 1. Dimensions of Data Quality

Domain	Dimension	Definition
Utility	Relevance	Relevance refers to whether the data product is targeted to meet current and prospective user needs.
	Accessibility	Accessibility relates to the ease with which data users can obtain an agency’s products and documentation in forms and formats that are understandable to data users.
	Timeliness	Timeliness is the length of time between the event or phenomenon the data describe and their availability.
	Punctuality	Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release.
	Granularity	Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any of a number of characteristics (e.g. demographic, socio-economic).
Objectivity	Accuracy and reliability	Accuracy measures the closeness of an estimate from a data product to its true value. Reliability, a related concept, characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.
	Coherence	Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data.
Integrity	Scientific integrity	Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence.
	Credibility	Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.
	Computer and physical security	Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.
	Confidentiality	Confidentiality refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party.

Although single source data collections have one set of quality threats, quality threats for integrated data products include the quality threats for each input source, threats resulting from integration methods, and when applicable, threats identified from additional statistical methods used to produce outputs (Brown *et al.* 2018, FCSM 2018). However, the net results on overall quality for integrated data may be best determined by aggregate assessment methods rather than by attempting to measure the cumulative consequences of each component and all possible interactions of components.

In addition, threats to data quality for the original use of a data source may differ from those for its secondary use or use in integrated products. While the importance of each threat to quality varies across data collections, data products, and analyses, the most significant threats to quality should be identified and minimized as resources permit as they are likely to affect the fitness for use across a wide range of objectives.



Best practices to identify threats to data quality include:

- Regularly identify threats to data quality for ongoing data collections, particularly when considering new source data for inclusion. Decisions on trade-offs among threats and mitigation measures should be considered in the context of the purpose of the data and all identified threats.
- For integrated data products, identify threats to data quality for all source data in the context of the purpose of the integrated data. The quality of a data source for use in an integrated data product may differ from that for its original purpose.
- For integrated data products, evaluate the quality of the integration method, including record linkage, modeling, and harmonization.
- For data outputs, including estimates in tables and reports, identify threats to data quality that arise from threats to the quality of the source data and from identified threats resulting from analysis.
- For some data outputs and integrated data products, measures of quality currently may be best identified by evaluations of critical processing and by evaluations of analysis decisions through sensitivity analysis, or by comparisons to benchmark or gold standard data.

The best level of detail for reporting data quality depends on the intended use and users of the information and the data product being documented. Quality reporting can take many forms, from detailed documentation about data collections to technical notes or footnotes accompanying a published statistic. Generally, reporting data quality for a particular data product or output will build on the documentation for the data collection from which it is derived and will include the documentation of additional quality assessments for any integration methods and statistical methods used to produce the output. Detailed technical documentation is needed within a data collection program for continuity planning and parts of the detailed technical documentation can be extracted for various data products to meet the needs of different types of data users (*e.g.* National Center for Educational Statistics (NCES) 2012). One example of this is the reuse of a subset of the content from the OMB Information Collection Request package to populate the Department's Data Inventory, a project that is under development at the Department of Education (OMB 2019). Differing amounts of technical detail can be tailored to meet the needs of different types of data users. High-level summaries encapsulating the key quality issues are particularly important. In a few sentences, such a summary would provide an overview of the data product's origin and describe its suitability for a particular use and the likelihood that key data outputs could lead to misleading information.



In summary, best practices for reporting on data quality include:

- Provide detailed descriptions of the significance of each potential threat for internal use, what countermeasures, if any, are taken, and what trade-offs or caveats are warranted. Citations to relevant studies that the agency or others have conducted to evaluate the threat and its likely impacts should be included. The material should be kept in a discoverable and accessible form for future data stewards, program managers, and analysts. Material should include all information required for Data Management Plans and metadata requirements.
- For users who require detailed information, including those who use data for secondary-products or perform data analysis, provide descriptions of any significant data quality threats. Technical documentation that accompanies the data product should summarize the likely consequences of identified threats, including the potential for the data to be unsuitable for a particular use.
- For occasional users of the data, including those who use tables, reports, and other data outputs for decision making, identify and summarize the most consequential quality threats that will help them to understand any limitations on the appropriate use of the data.



Future Directions

A wide variety of research is actively being conducted to identify new methods to meet emerging needs with the increasing use of data integration and secondary-use data to provide information for decision making. With integrated data, every phase of the process from data collection through dissemination is impacted. As agencies move these newly developed methods into production, often further research must be conducted to bridge the gap between a theoretical result and a method that will work well in a specific production environment. At the same time, new metrics of assessing data quality for statistics arising from integrated data are being developed. This document reflects the authors' best understanding of the current status of the work as it impacts agencies in their efforts to produce quality official statistics and future reports that build on the data quality framework will reflect the numerous advances anticipated as agencies gain more experience with integrated and secondary-use data.

Several topics and questions requiring additional study or research are identified in this report, including:

- Methods for assessing relevance and accessibility;
- Determinants of response, measurement error and coverage error in nonstatistical data;
- Methods for assessing the interaction between and among different types of quality threats;
- Methods for identifying, evaluating, and measuring cumulative errors for integrated data;
- Identifying characteristics of gold standard data for sensitivity analyses and another quality checks, including 'truth decks' for record linkage;
- Methods for assessment and protection of disclosure risk;

- Methods for communicating and reporting quality across various audiences using new technologies and methods; and
- Updated methods for creating templates and related tools for recording internal data quality documentation and converting that documentation into reports and data quality components of standard metadata that take advantage of new technologies (Statistical Community of Practice 2020).

Future evolution for identifying and reporting data quality will be positive if experiences, both successful and less successful, are channeled into learning agendas, as recommended by the Commission of Evidence-Based Policymaking (2017) and required by the Foundations for Evidence-Based Policymaking Act of 2018 (Pub. L. No. 115-435, 132 Stat. 5529, 2018). Through data producers and analysts sharing their experiences, the state-of-the-art will advance through the approaches outlined in this report and methods yet to be conceived. The FCSM stands ready to work with data stewards throughout the federal government to share experiences and move the state-of-the-art forward.

References

Biemer PP, de Leeuw E, Eckman S, Edwards B, Kreuter F, Lyberg LL, Tucker C, West BT (eds). 2017. Total Survey Error in Practice. Wiley Series in Survey Methodology. John Wiley & Sons, Inc. Hoboken, New Jersey.

Brown A, Abraham KG, Caporaso A, Kreuter F. 2018. Findings from the Integrated Data Workshops hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society, available at https://nces.ed.gov/fcsm/pdf/Workshop_Summary.pdf.

CNSTAT. National Academies of Sciences, Engineering, and Medicine, 2017. Innovations in federal statistics: Combining data sources while protecting privacy. National Academies Press.

CNSTAT. National Academies of Sciences, Engineering, and Medicine, 2018. Federal statistics, multiple data sources, and privacy protection: Next steps. National Academies Press.

Commission on Evidence-Based Policymaking. 2017. The Promise of Evidence-Based Policymaking, available at <https://www.cep.gov/cep-final-report.html>.

Czajka JL, Stange M. 2018. Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines. Washington, DC: Mathematica Policy Research, April 27, 2018.

FCSM. 2001. Measuring and reporting sources of error in surveys. Washington, DC: U.S. OMB (Statistical Policy Working Paper 31).

FCSM. 2018. Transparent Quality Reporting in the Integration of Multiple Data Sources: A Progress Report, 2017-2018. Federal Committee on Statistical Methodology. October 2018. Available at https://nces.ed.gov/fcsm/pdf/Quality_Integrated_Data.pdf.

ICSP. 2018. Principles for Modernizing Production of Federal Statistics, Available at <https://nces.ed.gov/fcsm/pdf/Principles.pdf>.

OMB. 2015. Memorandum M-15-15. Improving Statistical Activities through Interagency Collaboration, available at <https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2015/m-15-15.pdf>.

OMB. 2019c. "2020 Federal Data Strategy Action Plan", available at <https://strategy.data.gov/action-plan/>.

Pub. L. No. 106-554, § 515(a). 2000. Information Quality Act.

Pub. L. No. 115-435, 132 Stat. 5529, Foundations for Evidence-Based Policymaking Act of 2018.

Statistical Community of Practice (SCOPE) Metadata team. 2020. Metadata Systems for the U.S. Statistical Agencies, in Plain Language. Found at https://nces.ed.gov/fcsm/pdf/Metadata_projects_plain_US_federal_statistics.pdf.

Preface

The COVID-19 pandemic accelerated a transformation that was underway in the principal statistical agencies and among data programs throughout the federal government. Demands by decisionmakers and the public for daily information on the pandemic's effects placed a premium on new data sources that could deliver statistics in days and hours instead of months and years. The federal statistical system's traditional emphasis on carefully designed and disseminated data was overtaken by the quest for preliminary indicators that could be deployed quickly necessitating shortening some of the established quality-related tasks.

This transformation does not change the obligation for data program managers and analysts in all federal agencies to pay attention to data quality. Indeed, it is just as important now, if not even more important, to document data quality. Even if documentation is done after the fact, data users need to be able to understand the limitations and appropriate uses of the data. Data quality has many dimensions, such as accuracy and timeliness, that must be balanced to deliver answers to decision makers when the information is needed. As stewards of federal information, data program managers and analysts are responsible for understanding and explaining to the data user how the information can be used with confidence. They are also responsible for conducting ongoing data quality evaluations and releasing updated data products when warranted. While the COVID-19 pandemic increased the demand for timely data, it has also demonstrated that just providing the needed data is not sufficient. There remains a critical need for information about how the data were collected along with evaluations of the quality of the data for the different purposes for which they might be used. Addressing the quality metrics of any data collection system enhances the confidence of those analyzing the data and the recipients of the information products. The pandemic has shown that the need for confidence in data has not disappeared.

This report provides a framework for considering data quality issues and recommendations for documenting those considerations. Documentation is essential for the program managers and analysts who take over data products after the originators have moved on. For users of all types of data products, documentation is crucial to ensure that the quantitative information is not overextended or misused. The needed level of detail in documentation for program managers, "power users" of micro-data and data products, and the occasional user of data outputs differs substantially, as recognized in the conclusions of this report.

The authors of the report considered all kinds of data, including traditional sources as well as data originally collected for other purposes, such as administrative records and images collected from remote sensors. They recognize that many statistics originate from complex models that integrate data from multiple sources with both statistical and nonstatistical methods. The report provides a framework for identifying and documenting data quality issues across all types of data products, although the complexity of identification and reporting may be greater for new data products and for integrated data than for traditionally designed data products, such as surveys.

Considerations of data quality continue to evolve with new data sources and analytical methods. This report will be revisited as the state of the art and state of practice improve. The authors welcome feedback from the data community.

Acknowledgements

The FCSM Data Quality Working group thanks Brock Webb (Census Bureau) for his substantial input to the sections on computer and physical security and Barry Johnson (Statistics of Income Division, Internal Revenue Service and Interagency Council on Statistical Policy) for his guidance and encouragement throughout the project. The Working Group also appreciates the editorial and substantive contributions of FCSM members Jennifer Madans, Shelly Martinez, Jennifer Ortman, Anne Parker, and Robert Sivinski.

The FCSM Data Quality Workgroup acknowledges the participation and contributions of David Maron, statistician at the U.S. Department of Veterans Affairs and a member of this Working Group until his death on December 21, 2019.

List of Acronyms

AI	Artificial Intelligence
BEA	Bureau of Economic Analysis
BLS	Bureau of Labor Statistics
BTS	Bureau of Transportation Statistics
CIPSEA	Confidential Information Protection and Statistical Efficiency Act
CNSTAT	Committee on National Statistics
ERS	Economic Research Service
FCSM	Federal Committee on Statistical Methodology
FISMA	Federal Information Security Management Act
GDP	Gross Domestic Product
GSA	General Services Administration
HHS	Health and Human Services
ICSP	Interagency Council on Statistical Policy
IMF	International Monetary Fund
IT	Information Technology
NAICS	North American Industry Classification System
NASS	National Agricultural Statistics Service
NCES	National Center for Education Statistics
NCHS	National Center for Health Statistics
NCSES	National Center for Science and Engineering Statistics
NCVAS	National Center for Veterans Analysis and Statistics
NIST	National Institute of Standards and Technology
OECD	Organisation for Economic Co-operation and Development
OMB	Office of Management and Budget
SCOPE	Statistical Community of Practice
SOI	Statistics of Income
SSN	Social Security Number

1 Introduction

Federal data inform and improve our decisions, policies, and lives. It has been estimated government data guide trillions of dollars of investments and generate billions for the private sector, each year (Beede *et al.* 2015). Federal agencies disseminate a wide range of data products that are used pervasively by the public, businesses, and governments to inform critical decisions (OMB 2018). These data products are produced from a variety of data systems, collections, and processes, and include data files, online data tools for tabulations, analytic reports and other outputs. Traditionally, these products have been the result of statistical data collections designed for the purpose of creating statistics and other statistical products. Today, a major modernization effort within the federal statistical system is to acquire already collected program or administrative data from other parts of the government, or acquire data from nongovernmental sources, rather than initiating new, expensive and time-consuming data collections. As a result, data products increasingly include data originally collected for nonstatistical purposes as well as data sets that result from integrating one or more statistical and nonstatistical data products. Within the federal government, data are used in myriad ways (*e.g.*, operational uses, performance monitoring, program evaluation, and policy formation).

Federal data that are accessible, discoverable, and usable by the public have fueled entrepreneurship, innovation, and scientific discovery in a growing number of instances (OMB 2018). These advancements will continue to expand as innovative analytical tools are developed and data governance by federal agencies matures. Indeed, the Foundations for Evidence-Based Policymaking Act of 2018 (hereafter, the Evidence Act 2018) aims to advance the effectiveness of these data by better leveraging data as a strategic asset through improved governance, systematic planning of analyses, and increased sharing of valuable data assets.

Federal guidance to implement the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) emphasizes the concept of “fitness for purpose” as a touchstone for evaluating and communicating the quality of data made available to the public (OMB 2019a), noting that higher-impact uses of data require higher-quality data. This concept, also known as “fitness-for-use,” is sometimes treated as a broad definition of data quality. In a review of data quality in the international statistical community, Czajka and Stange (2018) note that the concept of data quality is almost universally associated with whether the data meets their intended purpose in operations, decision-making, and planning. In this view, data quality is best evaluated within the context of the intended uses of the data. However, data have always been used for purposes beyond those originally intended and opportunities for a wide range of secondary-use applications are increasingly being identified and implemented by federal agencies. Data quality should be described sufficiently to enable potential users to consider their fitness for their particular purposes. To do this requires a multidimensional approach, in which identifying and reporting on data quality entails answering a range of questions. For example: Are the data accurate and reliable? Do the data measure things of value at a useful level of detail? Can they be provided on time?

Increasingly, federal agencies are sharing data and acquiring data from sources external to the government for secondary use, including to create integrated data products. Data quality for secondary-use data, that is, data initially collected for one purpose but used for another, can be challenging to assess comprehensively. The statistical literature for official statistics is most extensively developed around data collected for a particular use; for instance, statistical data collected through sample surveys (FCSM 2001). But this traditional literature does not cover all issues that are important for secondary uses of data and for the variety of integrated data products using these sources. The statistical literature on the quality of integrated data continues to mature. Although methods developing for record linked and modeled data can have general implications,

many of the recent contributions address the measurement of data quality within specific contexts. Some of these lessons are analogous to the insights of the traditional, survey-based literature and some are entirely new. Unifying these ideas under a common framework is a way to facilitate their mutual development.

In this report, FCSM provides a comprehensive data quality framework to support data quality identification and reporting by federal agencies. The establishment of such a framework has been called for by the Committee on National Statistics (CNSTAT) (CNSTAT 2017, 2018). Further, the Interagency Council on Statistical Policy (ICSP) has stated that “agencies should work to adopt a common language and framework for reporting on the quality of data sets and derivative information they disseminate” (ICSP 2018). The framework establishes a common set of objectives for agencies to consider while managing data and provides a common language so that data quality reports and communications by different federal agencies may be more easily compared and synthesized. Some federal agencies have previously defined their own frameworks, and data quality principles, priorities, and approaches and it is relatively straightforward to map agency-specific concepts to the common framework. Since the framework elaborates on a broad definition of data quality to which all agencies must adhere, its application by federal agencies may expand their quality assessments to dimensions not previously addressed.

This report puts forth a set of best practices developed by the FCSM for data producers to use in identifying and reporting on data quality. Data producers, for consistency within this report, are the data stewards, program managers and analysts who collect, acquire, process, manage, analyze, and disseminate data for statistical uses. These best practices apply to internal documentation needed for program continuity and to the information provided to the various users of data products and outputs. As there are many types of data products from federal agencies, there is a wide variety of users, from researchers who use complex micro-data files to users of estimates disseminated in tables, reports, and other outputs for decision making, evaluation and research. The best practices in this report build on earlier FCSM reports focusing on the accuracy of survey data (FCSM 1988, 1990a, 1990b, 2001), extending these to all data as called for by the Federal Data Strategy (OMB 2019b, 2019c). The implementation of best practices includes assessment of the primary threats to data quality in each of its dimensions, where threats are factors that reduce the quality of the data.



Taking into account the variety of circumstances in which data are created and the variety of users who consume data quality reporting, the best practices describe the following actions:

- **Identification of threats to data quality.** Detailed analysis should be conducted to identify particular threats to data quality using general approaches for assessment that are applicable to statistical and nonstatistical data sources as well as integrated data.
- **Reporting threats to data quality.** Threats to each dimension of quality should be documented throughout the data lifecycle, as a part of normal business practice. Reporting should focus on the threats determined to have the most impact for a particular data product and its set of intended users, with potentially different levels of data quality reporting targeted to users with different needs. For integrated data, threats should be documented for input sources, integration methods, and outputs. However, reporting the net results on overall quality for integrated data may be best determined by aggregate assessment methods rather than by attempting to measure the cumulative consequences of each component and all possible interactions of components.

The FCSM expects to continue its leadership in data quality issues by building on the framework and best practices described in this report. The curation of data quality is at the core of FCSM's longstanding mission and its ongoing work. Previous FCSM reports describe best practices and have been considered authoritative sources for federal agencies and for the larger statistics profession. Due to their broad applicability, the data quality framework and best practices described herein should have resonance with these large communities. Although this report provides a milestone in efforts to update materials to incorporate a multidimensional concept of quality and extend it to nonstatistical and integrated data, it leaves several additional products to be pursued. These include: guidance for agencies to follow when acquiring data with high potential for secondary use (see M-19-15 (OMB 2019b)) and the Federal Data Strategy (OMB 2019)), recommended measures for documenting new challenges to data quality, and the development of new data quality initiatives to be considered as FCSM develops its plans for future research.

FCSM acknowledges that there are other useful approaches to organizing a data quality framework. In particular, other sources discuss data quality in the context of cycles by which data are collected, processed, and released. For example, OMB's Circular A-130 documents several stages through which information passes, including "collection, processing, dissemination, use, storage, and disposition, to include destruction and deletion" (OMB 2016, 29). In its 2017-18 sequence of workshops to gather insight on the data quality of integrated data products, the FCSM distinguished between three data life cycle stages (inputs, processing, and outputs) (Brown *et al.* 2018, FCSM 2018). The life cycle and data-quality-framework perspectives for reporting quality to the user are not mutually exclusive. In fact, it can be beneficial for agency documentation to use both perspectives, drawing upon mutually reinforcing terminology and concepts from each. On the one hand, it is possible that data producers and data users find a life cycle perspective more familiar or more natural than a hierarchical framework. If so, an agency can leverage that familiarity by discussing data quality in terms of the life cycle. On the other hand, one or more dimensions of the framework might be neglected in a life-cycle perspective. A benefit to an agency of referring to a unified, comprehensive framework is that it helps ensure that all the dimensions and details that can matter to a user are covered when reporting data quality.

The report is organized as follows. Chapter 2 describes the framework and defines its domains and dimensions. Chapter 3 describes threats to data quality for each of the dimensions within the framework, including mitigators and alignments and trade-offs among them, when applicable. Chapter 4 provides guidelines for best practices for identifying and reporting data quality and closes with some research questions and future directions. Appendix A provides additional background information on policies and earlier reports related to data quality used in the development of the framework. Appendix B provides more information on the dimension of accuracy and reliability for integrated data. This dimension remains a cornerstone of data quality for data products. Many well-developed and emerging statistical and survey methods have been developed to measure and improve accuracy and reliability.

Specific report content that defines data quality, describes the threats to data quality, and provides guidelines for best practices for identifying and reporting data quality, and directions for future research by domain and dimension when applicable can be easily found throughout the document by using the Framework For Data Quality Report index below.

Index

B

best practices

- identify threats to data quality 5
- reporting on data quality 6
- data producers 13
- credibility 41
- identifying and reporting data quality 48
- identifying threats to data quality include 49
- reporting data quality 50

D

data quality


- defining data quality 3
- components of data quality 3
- threats to data quality 3
- best practices for reporting on data quality 6
- defining data quality 17
- factors that affect data quality 30
- importance of reporting data quality 48
- identifying threats to data quality 49
- reporting data quality 50

T


threats

- data quality 3
- relevance 31
- accessibility 32
- timeliness 34
- punctuality 35
- granularity 36
- accuracy and reliability 37
- coherence 39
- scientific integrity 40
- credibility 41
- computer and physical security 43
- confidentiality 44
- data quality 49
- precision and sources of bias 68


Legend




Threats




Best practices




Moving forward




Data quality



Utility



Objectivity



Integrity

References

- Beede D, Powers R. 2015. Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data. 10.13140/RG.2.1.3354.8888.
- CNSTAT. National Academies of Sciences, Engineering, and Medicine. 2017. Innovations in federal statistics: Combining data sources while protecting privacy. National Academies Press.
- CNSTAT. National Academies of Sciences, Engineering, and Medicine, 2018. Federal statistics, multiple data sources, and privacy protection: Next steps. National Academies Press.
- Czajka JL, Stange M. 2018. Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines. Washington, DC: Mathematica Policy Research, April 27, 2018.
- FCSM. 1988. Quality in Establishment Surveys. Washington, DC: U.S. OMB (Statistical Policy Working Paper 15).
- FCSM. 1990a. Data Editing in Federal Statistical Agencies. Washington, DC: U.S. OMB (Statistical Policy Working Paper 18).
- FCSM. 1990b. Survey Coverage. Washington, DC: U.S. OMB (Statistical Policy Working Paper 17).
- FCSM. 2001. Measuring and reporting sources of error in surveys. Washington, DC: U.S. OMB (Statistical Policy Working Paper 31).
- FCSM. 2018. Transparent Quality Reporting in the Integration of Multiple Data Sources: A Progress Report, 2017-2018. Federal Committee on Statistical Methodology. October 2018. Available at https://nces.ed.gov/fcsm/pdf/Quality_Integrated_Data.pdf.
- ICSP. 2018. Principles for Modernizing Production of Federal Statistics, Available at <https://nces.ed.gov/fcsm/pdf/Principles.pdf>.
- OMB. 2018. "Statistical programs of the United States government." Annual Report.
- OMB. 2019a. Memorandum M-19-15. "Guidance on Improving Implementation of the Information Quality Act," available at <https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf>.
- OMB. 2019b. Memorandum M-19-18, "Federal Data Strategy - A Framework for Consistency," available at <https://www.whitehouse.gov/wp-content/uploads/2019/06/M-19-18.pdf>.
- OMB. 2019c. "2020 Federal Data Strategy Action Plan", available at <https://strategy.data.gov/action-plan/>.
- Pub. L. No. 106-554, § 515(a). 2000. Information Quality Act.
- Pub. L. No. 115-435, 132 Stat. 5529. 2018, Foundations for Evidence-Based Policymaking Act of 2018.

2 A Unified Data Quality Framework

Defining Data Quality

The FCSM Data Quality Framework presented in this chapter is broadly applicable to data produced by a range of processes and can be applied in a variety of settings. The framework readily applies to data collected for statistical purposes through sample surveys and censuses. The framework can also be applied to nonstatistical data that are collected by governments and businesses in the course of their administration of programs, data that are created from transactions, and data that are obtained by satellites and sensors. It applies to integrated data created using statistical data, nonstatistical data or both. And, it applies to data products such as estimates and other data outputs obtained using standardized methods, as well as to the data files that underlie such analyses. Although distinctions between data files and outputs, such as estimates, can be made, in practice some data products produced as outputs may be used as inputs to other data products through secondary use or data integration.

The framework provides a common foundation upon which federal agencies can make decisions about the management of data collections throughout their lifecycle, evaluating trade-offs among different quality dimensions where necessary. In particular, the framework provides a structure for identifying and addressing key data quality issues, applying accepted methods at an appropriate level of rigor, and for reporting on the quality of data products, supporting their effective use.

The framework is not novel—it synthesizes quality concepts currently employed by federal agencies, international organizations, and other authoritative bodies. A review of precedents to its development is given in Appendix A.

Data quality is the usefulness and credibility of data and products derived from data (e.g., statistics, analyses, and visualizations). Data and data products have high quality when they capture desired information using scientifically appropriate methods to represent reality in a manner that sustains public trust. This definition of data quality, informed by the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) and other sources, was developed by the FCSM for the purpose of the framework. The definition contains three broad components, highlighted in the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) and referenced in this framework as the *domains* of data quality: utility, objectivity, and integrity.¹

Definitions of the domains of data quality used in this framework are in Table 2.1. As with the definition of data quality, definitions of the components were developed for the framework by the FCSM using the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) and through careful review of documents and reports from national and international statistical agencies and organizations.

1. Utility, objectivity and integrity are also identified as the constituents of data quality in the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) and its associated OMB guidance (OMB Information Quality Guidelines, October 1, 2002). Although the FCSM Framework is intended to be broadly compatible with these sources, it should not be interpreted as specific guidance for Information Quality Act implementation.

Table 2.1. Domains of Data Quality

Domain	Definition
Utility	Utility refers to the extent to which information is well-targeted to identified and anticipated needs. It reflects the usefulness of the information to the intended users.
Objectivity	Objectivity refers to whether information is accurate, reliable, and unbiased, and is presented in an accurate, clear, and unbiased manner.
Integrity	Integrity refers to the maintenance of rigorous scientific standards and the protection of information from manipulation or influence as well as unauthorized access or revision.

The framework builds on these three domains, nesting 11 data quality *dimensions* within them, as shown in Figure 2.1 and Table 2.2. The dimensions represent areas in which specific aspects of data quality can be considered.

Figure 2.1. The FCSM Data Quality Framework

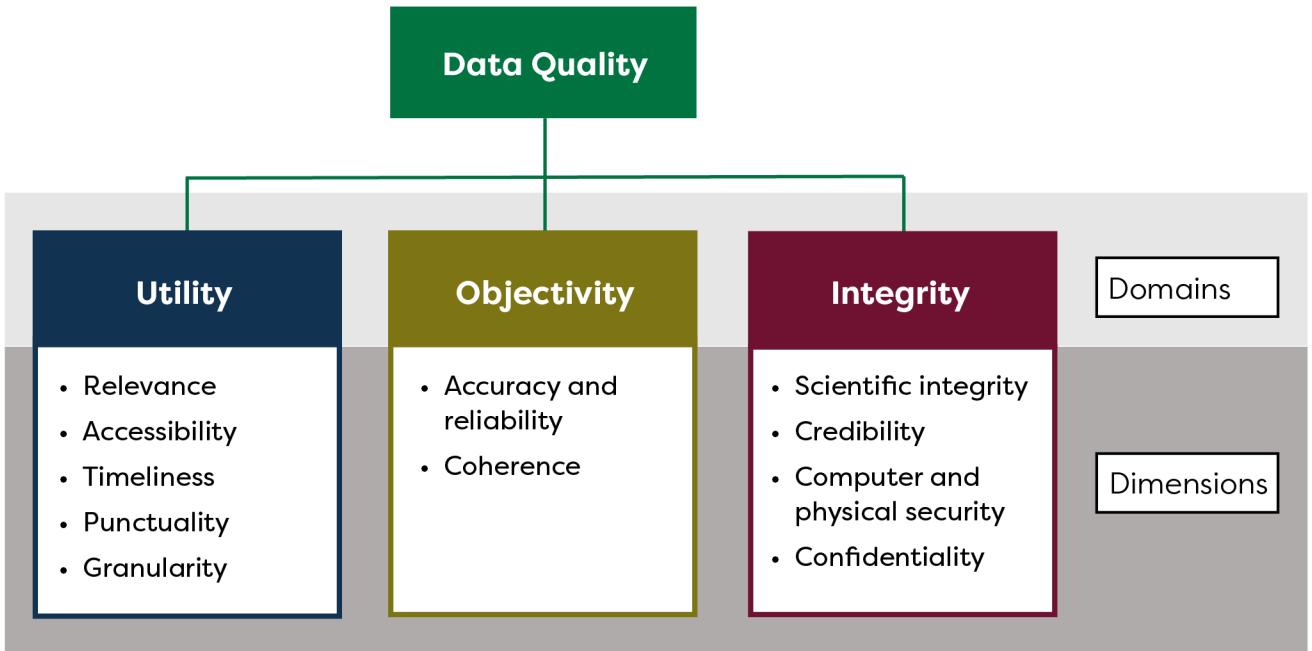


Table 2.2. Dimensions of Data Quality

Domain	Dimension	Definition
Utility	Relevance	Relevance refers to whether the data product is targeted to meet current and prospective user needs.
	Accessibility	Accessibility relates to the ease with which data users can obtain an agency’s products and documentation in forms and formats that are understandable to data users.
	Timeliness	Timeliness is the length of time between the event or phenomenon the data describe and their availability.
	Punctuality	Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release.
	Granularity	Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any of a number of characteristics (<i>e.g.</i> demographic, socio-economic).
Objectivity	Accuracy and reliability	Accuracy measures the closeness of an estimate from a data product to its true value. A related concept is reliability, which characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.
	Coherence	Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data.
Integrity	Scientific integrity	Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence.
	Credibility	Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.
	Computer and physical security	Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.
	Confidentiality	Confidentiality refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party.

To facilitate its conceptualization, discussion, and application, the framework has a hierarchical structure:

the three domains are depicted as distinct rather than overlapping, and each dimension is listed under the domain of its primary impact. This is an approximation. In practice, some dimensions may have implications for more than one domain. For example:

- Scientific integrity and credibility in a federal data program have a direct effect on the integrity domain but they also affect the utility of data products produced.
- Threats to a data product’s accuracy and reliability will directly impact its objectivity, but this will, in turn, affect the credibility, and thus the integrity, of its published estimates.

In this framework, the domains are essentially goals for data stewards, data program managers, and analysts, and the dimensions are the more specific objectives. An objective can serve more than one goal.

Utility

Utility refers to the extent to which the data product is well-targeted to identified and anticipated needs and reflects the usefulness of the data product to the intended users. High levels of utility result when agencies take potential uses of the information into consideration while designing measures for production. Utility is enhanced by continual assessment of information needs, the anticipation of emerging requirements, and the development of new data products and services to meet those needs and requirements (OMB 2006). Utility is related closely to the term “usefulness”—the extent to which the data meet the needs of their users. It encompasses whether the data are of interest, the ease by which users can access and use the data, and whether the data are credible with the users. Note that utility does not include how well the data approximate the intended indicator—such considerations are included in the objectivity domain.

As shown in Figure 2.1 and Table 2.2, the utility domain includes five quality dimensions: relevance, accessibility, timeliness, punctuality, and granularity. These are defined in the subsections below.

Relevance

Relevance refers to whether the data product is targeted to meet current and prospective user needs. Data are relevant when the outputs that are needed are produced and the outputs that are produced are needed (United Nations Economic Commission for Europe 2011, 38). Relevance is best achieved when the scope, coverage, reference period, geographic detail, data items, classifications, and statistical methodology meet user needs. Relevant data are aligned with the current, as well as any future needs, of users that may be anticipated. As summarized by Czajka and Stange (2018) in their review of international guidelines and standards for data quality reporting, alignment with the needs of users is often seen in definitions of relevance in many national statistical organizations, including Statistics Canada (Statistics Canada 2017), the European Statistical System (European Statistical System 2020), and the Australian Bureau of Statistics (Australian Bureau of Statistics 2009). The Evidence Act defines relevant statistical information as “processes, activities, and other such matters likely to be useful to policymakers and public and private sector data users” (Pub. L. No. 115-435, 132 Stat. 5529, 2018 codified at 44 USC 3563(d)(4)).

<p>Accessibility</p>	<p>Accessibility relates to the ease with which data users can obtain an agency’s products and documentation in forms and formats that are understandable to data users. Two main classes of activity support the accessibility of a data product: making the data product available to a broad range of users in easy-to-use formats and providing metadata and other documentation to facilitate the use and interpretation of the data (Statistical Community of Practice 2020). Documentation also provides users with information about the quality of the data.</p> <p>Data products have high availability when they are provided in discoverable, open (nonproprietary) formats that are accessible to users with vision or other impairments (ICSP 2018), and when users costs associated with access are limited (Statistics Canada 2000). Indeed, Title 2 of the Evidence Act requires public government data to be accessible (Pub. L. No. 115-435, 132 Stat. 5529, 2018, codified at 44 USC 3562 (d)(3)). Agencies can also enhance availability by using plain language geared to the intended audience to clearly and correctly present all information products (OMB 2006), and by ensuring that official statistics are disseminated in forms that enable and encourage analysis and reuse (United Kingdom Statistics Authority 2018).</p>
<p>Timeless</p>	<p>Timeliness is the length of time between the event or phenomenon the data describe and the availability of the data. Statistical Policy Directive No. 4, (OMB 2008, 12625) directs federal statistical agencies to “minimize the interval between the period to which the data refer and the date when the product is released to the public.” According to European Statistical System (2020), when a data collection requests data for a specified prior reference period, timeliness includes the period between the reference period and the date when the data are collected plus the production period from the end of data collection until the release of the first product.</p>
<p>Punctuality</p>	<p>Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release. This is important for meeting user expectations (especially legislated deadlines) and by precluding even the appearance of political interference in a scheduled release (OMB 1985, 2008). Punctuality can be expressed through a dichotomous measure of whether a data product was released on schedule, that is, the data release was either punctual or not (Czajka and Stange 2018).</p> <p>Punctuality encourages users to depend on data products because their timing can be anticipated and used for planning. Punctuality to the nearest second is important for data releases that move markets and are the subject of automated trading in the stock market. A performance measure for some agencies and data programs, <i>e.g.</i>, the National Agricultural Statistics Service (NASS), is the percentage of reports released on schedule.</p>

Granularity

Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time; level of geographic detail available; or the amount of detail available on any of a number of characteristics. This definition builds on CNSTAT (2018) where granularity is defined as “the degree to which estimates can be obtained for small subdivisions of the population, such as spatial subdivisions or different socioeconomic status categories.” Granularity is valuable to users when variations over space (*e.g.*, block level, town or city, county, state), time (*e.g.*, monthly, quarterly or annual), and other characteristics (*e.g.*, categories of race and ethnicity, gender identity, and socio-economic status of individuals; size and industry of establishments) are meaningful.

For example, consider the temporal granularity of travel. When averaged over a year, crowded holiday travel disappears, and when averaged over a day, rush hour congestion in urban areas is canceled by empty streets at night (Bureau of Transportation Statistics (BTS) 2019). Analogous considerations can affect analyses of phenomena that may vary over space or between groups: Do aggregate statistics for a state mask conditions of a city within the state? Do measures of health, education, and justice differ by race, ethnicity, age, gender, or socioeconomic status? Do annual estimates mask seasonal variation?

Objectivity

Objectivity refers to whether information is accurate, reliable, and unbiased, and is presented in an accurate, clear, and unbiased manner. Objectivity applies to both the substance of and the presentation of the information. Objectivity is achieved by using sound statistical and research methods to generate data and develop analytical results (OMB 2002, 8459).

The objectivity domain includes the data characteristics that are most traditionally and intuitively associated with data quality. Are the data based on sound methods of measurement that accurately capture the objects they claim to measure? Are the indicators/estimates presented in an accurate way? These two elements—substance and presentation—are described in OMB’s definition of objectivity. The Evidence Act reinforces this definition, referring to objective statistical activities as those that are “accurate, clear, complete, and unbiased.” Note that some sources include the provision of documentation as part of objectivity (see, *e.g.*, OMB 2014, 71615). However, under this framework, documentation is primarily a contributor to accessibility—a dimension of the utility domain—as it relates directly to the users’ experience with the data that are produced, rather than the production of the data itself.

As shown in Figure 2.2 and Table 2.2, the objectivity domain comprises two quality dimensions: a) accuracy and reliability, and b) coherence. These are defined in the subsections below.

Accuracy and reliability

Accuracy measures the closeness of an estimate from a data product to its true value. A related concept is reliability, which characterizes repeated estimates of accuracy over time. For an estimate to be accurate, all components of the data product need to be accurate. While accuracy can apply to an entire data file or data collection, many common measures, including standard errors and other measures of precision, apply to estimates. Accuracy for outputs of integrated data depends on the accuracy of source data and linkage or modeling errors that result from the integration process. Reliability, a related concept, characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.

The concepts of accuracy and reliability are closely related. The Evidence Act defines accurate statistics as those that “consistently match the events and trends being measured” (Pub. L. No. 115-435, 132 Stat. 5529, 2018, codified at 44 USC 3563 (d)(1)). Some sources simply cite the concepts in tandem, Czajka and Stange (2018, vix) state that “accuracy and reliability refer to the degree to which statistical information correctly describes the phenomena it was designed to measure.” A tandem characterization is also known as construct validity in the evaluation literature (Cook and Campbell 1979, Shadish *et al.* 2001).

Coherence

Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data. Coherence applies to data over time, across key domains and when data originate from different internal and external sources. There are many definitions of coherence used to describe data. The definition of coherence for the framework is similar to that used in FCSM (2001) from Depoutot and Arondel (1999): “the ability of the statistical data program to maintain common definitions, classifications, and methodological standards when data originate from several sources.” Comparability of statistics, which the FCSM defines as “the ability to make reliable comparisons over time,” is a subset of coherence (FCSM 2001). According to Statistical Policy Directive 2, “(a) consistent data series maintains comparability over time by keeping an item fixed, or by incorporating appropriate adjustment methods in the event an item is changed.” (OMB 2006, 29) The European Statistical System (2020) defines coherence and comparability as the “adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics.” However, coherence should not be understood as the maintenance of data series when change in methods are required in response to changing data needs or changes in the concepts being measured that reflect societal change. When change is required, coherence requires that appropriate methods to bridge across the change be developed and documented.

For the framework, coherent data products, whether from national surveys, statistical and non-statistical sources or integrated products, are consistent and comparable with other relevant data sources of known and documented quality, across key subdomains, and over time.

Integrity

Integrity refers to the maintenance of rigorous scientific standards and the protection of information from manipulation or influence as well as unauthorized access or revision. Integrity is a domain that captures whether data products are produced and managed appropriately. The integrity of a data product can be affected by the attributes of the data producer. Promoting integrity entails safeguarding data from improper use, whether through manipulation of handling, estimation and dissemination processes, unauthorized access, or re-identification of confidential data elements. The Information Quality Act defines integrity as “the protection of information from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification.” (Pub. L. 106-554, 2000) Others identify integrity with impartiality (United Kingdom Statistics Authority 2018) and scientific practice (OECD 2012). Our definition includes both of these aspects, as well as confidentiality protection.

As shown in Figure 2.1 and Table 2.2, the integrity domain includes four quality dimensions: scientific integrity, credibility, computer and physical security, and confidentiality. These are defined in the subsections below.

Scientific integrity

Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence. The “Statement of Commitment to Scientific Integrity by Principal Statistical Agencies” (Principal Statistical Agencies 2012) supports the breadth of factors that contribute to the integrity of official data products:

“Methodological improvements and rigorous approaches to data collection and analysis require the application of scientific methods. Computer scientists, demographers, economists, geographers, mathematicians, survey statisticians, and other scientists are needed for producing high quality, objective statistics from surveys or administrative data. Subject area experts, such as epidemiologists and engineers, are also needed to maximize data quality. Research and methodological innovation are required to continuously improve the quality and scope of our data products while protecting privacy and ensuring confidentiality. All of the above mentioned factors are critically important to ensuring the credibility of Federal statistical agencies” (Principal Statistical Agencies 2012).

For related perspectives on scientific integrity for official statistics, see the Office of Science and Technology Policy Memorandum on Scientific Integrity (Holdren 2010, Government Accountability Office 2019), the IMF (2003) “Assurances of Integrity,” and the United Kingdom’s *Code of Practice for Statistics* (United Kingdom Statistics Authority 2018).

Credibility

Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer. In other words, credibility is an attribute of a data producer that is attached to the products that the producer disseminates. Perceptions of data users about the reputations of data producers and their products have a direct impact on the integrity of the data product. An important aspect of credibility is “trust in the objectivity of the data.” (Czajka and Stange 2018, 55)

OMB Statistical Policy Directive 1: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units (Directive 1) highlights the essential role that credibility plays in the utility of data disseminations (OMB 2014, 71610):

“The Nation relies on the flow of credible statistics to support the decisions of individuals, households, governments, businesses, and other organizations. Any loss of trust in the relevance, accuracy, objectivity, or integrity of the Federal statistical system and its products can foster uncertainty about the validity of measures our Nation uses to monitor and assess performance, progress, and needs.”

Although Directive 1 establishes credibility as a fundamental responsibility of statistical agencies and units, it is a goal for any agency disseminating data.

Computer and physical security

Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification. This definition is based on prior OMB guidance (OMB 2002, 2006) and the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000). The Information Quality Act indicates that agencies may rely on their implementation of the federal government’s computer security laws “to establish appropriate security safeguards for ensuring the integrity of the information that the agencies disseminate” (OMB 2002). Similarly, the United Nations also requires an IT-security policy to be in place “for the protection and security of personal data” and that the IT-security policy must ensure that “Statistical agencies secure adequate survey methods and processing methods and guarantee that data are not falsified by human or technical misbehaviour.” (United Nations 2015, 57).

Physical security can complement computer security. The European Commission (2011) requires that “Physical, technological and organizational provisions are in place to protect the security and integrity of statistical databases.” In the United States, OMB guidance requires that “agencies must implement safeguards throughout the production process to ensure that survey data are handled to avoid disclosure. . .”

Confidentiality

Confidentiality refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party. This definition is from the Evidence Act (Pub. L. No. 115-435, 132 Stat. 5529, 2018, codified at 44 USC 3563(d)(4)) and is consistent with requirements that individually-identifiable data be protected from unauthorized disclosure in order to uphold its confidentiality (OMB 2006), in which individually-identifiable data are those that permit the identity of the respondent or entity to which the information apply to be reasonably inferred by either direct or indirect means.

Many international statistical agencies and organizations also identify confidentiality as a key value. For example, the United Kingdom's Statistical Authority (2018 9) states that "Private information about individual persons (including bodies corporate) compiled in the production of official statistics is confidential and should be used for statistical purposes only." The European Commission (2011) acknowledges the privacy of data providers (*i.e.*, households, enterprises, administrations, and other respondents) and expresses a commitment to protect the confidentiality of the information respondents provide and guarantee its use only for statistical purposes. The United Nations Statistical Commission (1994) states that "(i) individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes."

References

Australian Bureau of Statistics. 2009. The Australian Bureau of Statistics Data Quality Framework. Canberra. Available at <https://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0>.

BTS. 2019. Transportation Statistics Annual Report. Available at <https://www.bts.gov/TSAR>.

CNSTAT. National Academies of Sciences, Engineering, and Medicine. 2017. Innovations in federal statistics: Combining data sources while protecting privacy. National Academies Press. Available at <http://www.bing.com/search?q=cnstat+statistics%3A+Combining+data+sources+while+protecting+privacy&src=IE-SearchBox&FORM=IESR4A>.

CNSTAT. National Academies of Sciences, Engineering, and Medicine. 2018. Federal statistics, multiple data sources, and privacy protection: Next steps. National Academies Press. Available at <https://www.nap.edu/catalog/24893/federal-statistics-multiple-data-sources-and-privacy-protection-next-steps>.

Cook TD, Campbell DT. 1979. Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Czajka JL, Stange M. 2018. Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines. Washington, DC: Mathematica Policy Research, April 27, 2018. Available at <https://www.mathematica.org/our-publications-and-findings/publications/transparency-in-the-reporting-of-quality-for-integrated-data-a-review-of-international-standards>.

Depoutot R, Philippe A. 1999. International Comparability and Quality of Statistics. In: Biffignandi S. (eds) Micro- and Macrodata of Firms. Contributions to Statistics. Physica-Verlag HD.

European Commission. 2011. European Statistics Code of Practice. Brussels, Belgium: European Statistical System Committee, September 28, 2011. Available at: <https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice>.

European Statistical System. European Statistical System Handbook for Quality Reports, 2014 edition. 2015. Luxembourg: Publications Office of the European Union. Available at <http://ec.europa.eu/eurostat/documents/3859598/6651706/KSGQ-15-003-EN-N.pdf/18dd4bf0-8de6-4f3f-9adb-fab92db1a568>.

Eurostat. 2020. European Statistical System Handbook for Quality and Metadata Reports, 2020 edition. 2020. Luxembourg: Publications Office of the European Union. Available at <https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf/bf98fd32-f17c-31e2-8c7f-ad41eca91783>.

FCSM. 2001. Measuring and reporting sources of error in surveys. Washington, DC: U.S. OMB (Statistical Policy Working Paper 31). Available at <https://nces.ed.gov/FCSM/pdf/spwp31.pdf>.

Government Accountability Office. 2019. Scientific Integrity Policies: Additional Actions Could Strengthen Integrity of Federal Research, GAO-19-265, April 2019. Available at <https://www.gao.gov/assets/700/698231.pdf>.

Holdren JP. 2010. Memorandum for the Heads of Executive Departments and Agencies. Subject: Scientific Integrity. December 17, 2010. Available at <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/scientific-integrity-memo-12172010.pdf>.

ICSP. 2018. Principles for Modernizing Production of Federal Statistics. Available at <https://nces.ed.gov/fcsm/pdf/Principles.pdf>.

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

OMB. 2002. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies, 67 FR 8451.

OMB. 2006. Standards and Guidelines for Statistical Surveys. September 2006. Available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

OMB. 2008. Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies (73 FR 12621, March 7, 2008)

OMB. 2014. Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units (79 FR 71609, Dec 2, 2014)

Organisation for Economic Co-operation and Development (OECD). 2012. Quality Framework and Guidelines for OECD Statistical Activities. Version 2011/1. Paris: OECD, 2012. <http://www.oecd.org/dataoecd/26/38/21687665.pdf>.

Principal Statistical Agencies. 2012. Statement of Commitment to Scientific Integrity (last modified on May 23, 2012). Accessed on June 24, 2020 from: <https://www.bls.gov/bls/integrity.htm>

Pub. L. No. 106-554, § 515(a). 2000. Information Quality Act.

Pub. L. No. 115-435, 132 Stat. 5529. 2018. Foundations for Evidence-Based Policymaking Act of 2018.

Statistical Community of Practice (SCOPE) Metadata team. 2020. Metadata Systems for the U.S. Statistical Agencies, in Plain Language. Found at https://nces.ed.gov/fcsm/pdf/Metadata_projects_plain_US_federal_statistics.pdf.

Shadish WR, T. D. Cook TD, Campbell DT. 2001. Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Boston: Houghton Mifflin.

Statistics Canada. 2000. The Standards and Guidelines on the Documentation of Data Quality and Methodology, March 30, 2000. Available at <https://www.statcan.gc.ca/eng/about/policy/info-user>.

Statistics Canada. 2017. Statistics Canada's Quality Assurance Framework. 3rd edition. Ottawa: Statistics Canada. Available at <https://www150.statcan.gc.ca/n1/en/catalogue/12-586-X>.

United Kingdom Statistics Authority. 2018. Code of Practice for Statistics: Ensuring Official Statistics Serve the Public. Edition 2.0. February 2018.

United Nations Economic Commission for Europe (UNECE). 2011. Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices. Geneva, Switzerland: United Nations Publication.

United Nations. 2015. United Nations Fundamental Principles for Official Statistics Implementation Guidelines. United Nations, January 2015. Available at https://unstats.un.org/unsd/dnss/gp/Implementation_Guidelines_FINAL_without_edit.pdf.

3 Factors that Affect Data Quality

This chapter provides information on factors that affect each of the 11 data quality dimensions and lists specific threats that can detract from the quality of a data product along each dimension. Alignment of similar and opposing threats across dimensions and potential trade-offs among them are described. No effort is made to recommend or suggest approaches to address each of the threats, although some approaches to limit or mitigate them and to measure their impacts are discussed, particularly those threats and mitigations that affect another dimension of data quality. To give a sense of the quality landscape, an effort has been made to draw attention to many of the salient data quality considerations and threats faced by data producers. However, the discussion is not exhaustive. Future FCSM reports and case studies may build on the framework to provide additional insights.

Utility

As described in Chapter 2, data products have high utility when they target the informational needs of key users, elicit the trust of those users, are provided in a way that users can easily access and apply, are recent enough to be actionable, are provided on schedule, and have an appropriate level of detail. These factors determine their potential value to the user for given levels of other dimensions, particularly accuracy. Note that a high-utility data product can have significant value to some users even if its accuracy is limited; however, the limitations need to be taken into consideration and documented. If a data product answers a key question for which no other information is available, it may be usefully applied to make key decisions when the limitations of the data do not have a major impact on how the data are used. By the same token, a highly accurate data product may have little value to users if they do not need the information, if they have alternative ways of getting it, or if it is provided too late. The different uses that a data set might be put to requires that utility be evaluated according in terms of multiple uses. This requirement complicates the assessment of utility and requires agencies to have a good understanding of user needs.

To be relevant, a data product's scope, coverage, reference period, geographic detail, data items, classifications, and methodology must meet user needs. Relevance is threatened when agencies do not understand the needs of users and when other data products fill those needs. Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) guidelines support utility by stressing the need for each agency to take public uses of information into consideration along with the agency's own intended uses (OMB 2002, 8451). OMB gives more specific guidance for survey data (OMB 2006), which is broadly relevant for all disseminated data products. In addition, the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) identifies several steps that federal agencies should take to ensure that disseminated information meets the needs of the intended users, including:

- Conduct internal reviews, analyses, and evaluations
- Request feedback from advisory committees, researchers, policymakers and the public
- Identify the audience for each information product

Agencies gather information about the relevance of their data products to users through a variety of tracking measures, *e.g.*, the number of downloads and frequency of citations in scientific literature and the popular media. In addition, they gather feedback through a variety of means (*e.g.*, environmental scans, survey-based tests, and expert panels) to anticipate changes in user needs and identify unmet needs that have already emerged. Internal reviews and analyses can also provide some sense of the usefulness of a data product. For example, the passage of the Affordable Care Act prompted

Relevance

agencies to conduct studies to identify relevant gaps in the available data for health insurance coverage (see, *e.g.*, Health and Human Services (HHS) Data Council 2011, Rabe *et al.* 2016, Schildkraut *et al.* 2015). But since data products are used as one of many inputs to a broad range of production and decision-making processes, direct measures of a data product's ultimate value can be elusive.



Threats to Relevance

1. Difficulties in thoroughly understanding and aligning user needs (*i.e.*, requirements). Some data may be misaligned with user needs if those needs are not readily apparent or were not considered in the collection, processing, or analysis of the data. As the needs of users are diverse, meeting the needs of some users may reduce the ability to meet the needs of others. Data documentation that clearly states appropriate uses can help users determine whether the data are aligned with their needs.
2. The availability of related data products. If similar information on a particular subject is readily available through alternative channels, a given data product may not be critically important. Consequently, changes in the availability of other sources of data (*i.e.*, new and emerging sources) can create a need for data producers to re-evaluate the relevance of a data product. For example, as price indices derived from private sources have grown in feasibility and use, the Bureau of Labor Statistics (BLS) has engaged with the developers of these private indices to ensure its understanding of how they affect the value of BLS's price indices. This threat is related to similar threats to the dimension coherence, which is also affected by the availability of related data products.
3. Negative perceptions of users. Subjective perceptions and understandings affect relevance, especially if users do not understand the applicability of a given data product to their needs. This threat is also associated with the scientific integrity and credibility dimensions in the domain integrity, through the threats of obsolescence and political interference.

Relevance

Accessibility relates to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable and interpretable to data users. To be accessible, data must be easy to find (*i.e.*, discoverable), easy to obtain at little or no cost using commonly available formats, and understandable. Data and supporting documentation must also be preserved through good archival practice to be accessible in the future. Providing metadata facilitates the use and interpretation of data, and increases their accessibility (Statistical Community of Practice 2020). Accessibility is threatened by high costs to access data or documentation, when disclosure limitation methods are applied, and when data products are confused with others in web-based searches for information.

Accessibility

One example of a successful effort to improve the accessibility of a large dataset is NASA's Big Earth Data Initiative. One focus of this Initiative was to standardize the formats, interfaces and protocols of federally funded, earth-observing data, using community input on the standards, in order to increase the interoperability of such data. The Initiative enhanced discoverability of these data by ensuring that NASA's earth-observing data were complete, searchable, and conformed to international standards. The NASA Big Earth Data Initiative also pursued other means to increase accessibility to these data, such as providing services to enable greater use and increasing the speed by which earth-observing data could be delivered for time-critical purposes, such as wildfire assessments (Blumenfeld 2016).

High-quality documentation provides contextual information about the contents, attributes, and methods of access of a data product (ICSP 2018), meets the needs for program continuity and dissemination of data files, and allows users to correctly interpret the data (FCSM 2001). Interpretability is supported when data are accompanied by “full and frank commentary” (United Kingdom Statistics Authority 2018). Documentation should contain information about data quality, particularly identifying threats to data quality with the most impact, enabling fitness-for-use assessments (Statistics Canada 2000).

High-quality documentation facilitates reproducibility, the ability to obtain consistent results from the same input data. Reproducibility is a key feature of high-quality data that supports credibility among users and allows improvements in data processes. CNSTAT recommends several steps for improving reproducibility (CNSTAT 2019a, 2019b) and urges agencies to fully document the processes used to produce and disseminate statistical products, and to take proactive steps to preserve data for future use. This documentation should include concepts, definitions, data collection methods, and describe factors that affect data quality.

OMB guidance emphasizes the need for documentation at every stage of the data lifecycle (OMB 2006). Internal documentation can be used to tailor additional reports on various data products, including data files and data outputs. Documentation of data quality for integrated data will include reference to documentation available for each input data source, documentation of integration and other statistical methods employed, and documentation of any outputs obtained.

Finally, high-quality data documentation is also important for supporting subject matter and technical research that may expand the usability of the data. Such documentation may include consultation, training, and technical assistance in addition to the contextual and process information described above.

Accessibility



Threats to Accessibility

1. Costs to access data. Open data requirements of the Evidence Act require that all government data be made available for little or no cost. However, an agency’s promise of confidentiality places restrictions on how data are accessed that can include costs associated with access. Data users may gain access to restricted data through special arrangements, such as those provided by Federal Statistical Research Data Centers—facilities for access to restricted-use micro data (Census Bureau 2020)—but such arrangements typically require time and expense to utilize. While costs are incurred, without access through Research Data Centers, access would not be available without threatening confidentiality, a dimension in the integrity domain.
2. Use of disclosure limitation methods. A variety of techniques are available to mask individual responses when making confidential data available to the public (see, *e.g.*, Abowd and Schmutte 2019). Although disclosure limitation methods provide a way to maintain confidentiality and allow data access in support of the quality principles related to these characteristics, the methods affect accuracy and reliability, a dimension in the objectivity domain, and may affect dimensions timeliness and granularity in the utility domain.

Accessibility

3. Costs to create effective documentation. Maintaining and curating metadata and technical documentation and communicating key data details to users using appropriately targeted language is resource intensive. The issue is complicated by the diversity of needs and preferences among the users of a given data product, as well as the breadth of potential issues affecting the data quality. Additionally, changing technologies may cause preferred modes of dissemination to change over time and hinder the ability to compare current with prior data product releases. Creating detailed documentation can impact timeliness, another dimension in the utility domain.
4. Confusion with other data products. Discoverability of a data product may be challenged when data catalogs and web-based searches include the data product, but it is difficult to identify in a mass of other, similar, entries. At the same time, when a data product does not use a standardized set of formats and metadata standards, potential users may discover only some subsets of the data product without recognizing other subsets.

Timeliness is the length of time between the event or phenomenon the data describe and their availability. Data producers consider tradeoffs between timeliness and other quality dimensions when determining what data products to create and how to create and document them. Timeliness is threatened by the availability of essential source data needed for the data products, processing time, the application of high levels of statistical rigor, and the preparation of documentation. Threats to timeliness are greater for phenomena and data products that change over time compared to those that change little over time.

Approaches for improving timeliness depend on the data product and the threat. Some approaches may include the release of provisional data, the use of models to estimate current conditions based on stable trends, by use of early returns from key respondents, and by the use of surrogate measures. One example of an agency's approach to balancing competing priorities between timeliness and other quality dimensions is demonstrated by the "release cycle" the Bureau of Economic Analysis (BEA) uses to disseminate estimates of Gross Domestic Product (GDP). Successively updated vintages of GDP estimates exemplify increasing accuracy as timeliness decreases (see the discussion in Prell *et al.* 2019, 29). Another approach that some agencies are considering is the use of forecasting models to generate predicted values of measures whose true values are only known with a lag. Such "nowcasting" models may exploit auxiliary information to anticipate movements in the official measure (see, *e.g.*, Cajner *et al.* 2019, Glaeser *et al.* 2019, and NCHS 2020a). Similarly, the NASS uses statistical models to combine its survey data on crop yields with administrative and weather data to produce forecasts for the primary crops (see, *e.g.*, Adrian 2012, Cruze 2016, Cruze and Benecha 2017).

BTS has used a variety of techniques to improve the timeliness of monthly airline passenger counts, which normally take a month for airlines to report and another month for BTS to process and resolve inconsistencies in the reports, BTS had used historical trends for preliminary estimates of airline travel. In 2020, in response to the COVID-19 pandemic, BTS shifted to preliminary estimates of change based on reports from the biggest airlines, which tend to report more quickly than the smaller carriers. In response to requests for faster updates, BTS began to use daily counts of airport security screenings as a proxy for airline travelers, recognizing that screenings are a rough indicator since they include airline and airport employees as well as passengers (BTS 2020).

Timeliness



Threats to Timeliness

1. Significant lags of one or more sources of input data. For data products comprising multiple sources, one or more data sources may only be available with a significant lag. To mitigate this threat, model-based methods, preliminary data, and proxies can be used. However, these approaches will increase the time needed for documentation and may increase threats to accuracy and reliability in the objectivity domain.
2. Processing time needed for appropriate use of source data. Some source data may entail significant work to address potential measurement and representation errors. The more processing that source data require, the less timely data products will be. Although federal agencies may be able to speed up such processing, increased timeliness may be costly in terms of both resources and accuracy, including error assessments and mitigations.
3. Statistical and methodological rigor. Applying rigor to the process of producing data products is time consuming. The magnitude of this threat increases with the complexity of the data product. Evaluations of linkage error and sensitivity analyses for assessing modeling assumptions are time consuming to conduct and to document. Federal agencies seek to find the appropriate balance in terms of accuracy, cost, and speed when applying statistical and methodological rigor to produce data products. To mitigate threats to timeliness, less statistically rigorous approaches may be employed. However, applying less rigorous methods may increase threats to accuracy and reliability in the objectivity dimension.
4. Production of effective documentation. Documenting data products is time consuming. This threat to timeliness increases with the complexity of the data product (*e.g.*, development and evaluation of integrated and model-based products), when producing documentation for new products, and for documenting using new platforms or processes. Although documentation is a threat to timeliness, the availability of comprehensive documentation can improve timeliness of subsequent releases of the same product for data producers and it increases accessibility and credibility, which are dimensions in the utility domain.

Timeliness

Punctuality

Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release. It can be expressed through a dichotomous measure of whether a data product was released on schedule; that is, the data release was either punctual or not (Czajka and Stange 2018). Punctuality is an expectation that increases utility and integrity with a transparent schedule that conveys that data releases will not be delayed, including for political reasons. Data products not released on schedule or with large time lags from planned target date to release increase threats to other dimensions of utility, including relevance and timeliness, and increase threats to the dimension credibility in the integrity domain. Low response or participation rates, external events, changes in priorities, and changes in the availability and content of external data sources can threaten punctuality. Some mitigators of threats to punctuality, such as reducing time spent on processing checks, sensitivity analysis, or documentation, may increase threats to other quality dimensions, including accuracy and reliability or accessibility.

Statistical Policy Directives No. 3 and No. 4 (OMB 1985, 2008) call for federal statistical agencies to annually publish the release dates for regular and recurring reports for the upcoming year, indicating

when each data product is expected to be released during the upcoming calendar year. Data identified by OMB as principal federal economic indicators must include an announcement of the next release date and time in each publication. In addition, when data for a series are released quarterly or more frequently, the time between the close of the reference period and the public release date should be at most 22 working days (OMB 1985). A comparison of these dates with the actual release dates allows the public to monitor the punctuality of data product releases.



Threats to Punctuality

1. Low response and participation rates. Low response or participation rates that result in longer data collection periods can affect the ability of the data product to be released on schedule.
2. External events. Unforeseen external events, such as a federal government shutdown or the coronavirus pandemic, disrupt the operations of affected agencies.
3. Changes in secondary-use source data. Changes to the collection, production, and availability of data acquired for secondary use are often outside of the control of agencies using data from external sources.
4. Changes in agency priorities. Changing priorities and resources as a result of changes in leadership or other factors may affect punctuality.

Punctuality

Granularity refers to the amount of disaggregation available for data products. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any number of demographic or socio-economic characteristics.

The granularity that agencies can feasibly produce is often limited by a combination of sample size constraints and confidentiality concerns. Data collection budgets often determine sample size tradeoffs that affect the granularity of the data collected and released. In order to protect the confidentiality of individual members of a dataset, agencies must draw from a substantial number of observations to produce each statistic. Because in many cases behaviors and policy effects vary among narrowly-defined groups, data users are often interested in greater granularity than agencies can offer.

Granularity

Concatenating multiple data sources over time can potentially improve granularity by increasing the number of observations within subsamples of interest if key estimates are thought to be stable over time. Identifying alternative data sources for secondary use or integration can mitigate some threats by providing more detail for subgroups. However, integrating data sources can also diminish the granularity if data require coarsening for harmonization or to align coverage.

The granularity in a data collection or restricted-access data file may differ from the granularity in publicly released data. To balance confidentiality and granularity, agencies may choose to use disclosure avoidance techniques or establish special user access provisions as highlighted under accessibility. Alternatively, some statistical treatments, such as perturbations, may be used to introduce artificial noise into data files and/or estimates, but such treatments come at the expense of accuracy and reliability, a dimension in the objectivity domain, and must be documented.

Granularity



Threats to Granularity

1. Small sample size. Smaller sample sizes decrease the statistical precision of data outputs and reduce the granularity for stable estimates. Sample size also affects similar threats to precision in the dimension accuracy and reliability.
2. Unavailable data. Data users are often interested in greater granularity than can be obtained. For surveys, underlying population sizes for some subgroups may be too small to be sampled efficiently. For data acquired for secondary use, such as administrative data or satellite data, the desired detail for the statistical use may not have been needed or collected for the original purpose of the data.
3. Confidentiality protections. Disclosure risks increase with granularity. Data for small population groups, low population geographic units, and narrow time intervals increase the risk of identifying respondents or confidential information from administrative records or other sources. Use of perturbation or other disclosure protections to ensure confidentiality affects other data quality dimensions, including accessibility in the utility domain and accuracy and reliability in the objectivity domain.

Objectivity

Objectivity has been the traditional focus of data quality. Are the data products accurate and reliable? Do they measure what they are intended to measure? Are they consistent over time? Are they consistent with other products produced from related data sources? Prior FCSM reports have focused on these areas, including Working Paper 31 “Measuring and Reporting Error in Surveys” (FCSM 2001). This report considered several concepts of objectivity related to data quality but focused on the measurement and reporting of various error sources that affect data quality: sampling error, nonresponse error, coverage error, measurement error, and processing error.

Accuracy and reliability

Accuracy measures the closeness of an estimate, including direct and modeled estimates, to its true value (FCSM 2001, Eurostat 2020). Accuracy of an estimate is dependent on the accuracy of its components. While some measures of accuracy apply to an entire data file or data collection (*e.g.*, response rate), standard errors and other measures of precision apply to estimates. Accuracy for outputs of integrated data depends on the accuracy of source data and linkage or modeling errors that result from the integration process. Reliability, a related concept, characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.

Threats to accuracy and reliability are briefly described below, drawn, in part, from the terminology of the Total Survey Error paradigm (*e.g.*, Biemer *et al.* 2017, Groves *et al.* 2009) and building from FCSM (2001), including sampling error, coverage error, measurement error and nonresponse error. Additional threats are modeling error, linkage error, and harmonization error that can affect a variety of data, but particularly integrated data. More details and discussions about threats to accuracy and reliability for integrated data are provided in Appendix B. Most threats to accuracy and reliability affect threats to many other dimensions, including relevance and granularity in the utility domain, coherence in objectivity domain objectivity, and credibility, confidentiality and scientific integrity in the integrity domain.

Accuracy and reliability

Threats to accuracy can be described as various “types” or “sources” of errors, which can reduce accuracy by diminishing either statistical unbiasedness, precision, or both. The accuracy of a data product reflects the accuracy of the input sources and all processing and calculations performed to transform those data into outputs. As these errors can accumulate throughout the data lifecycle, the accuracy of a data product will reflect a combination of the accuracy of its input sources, the processing steps applied to those inputs, and any additional calculations performed to transform the data into outputs. Data producers identify and report on the impact of many factors as a normal course of business. For example, OMB standards for survey data (OMB 2006) discuss the importance of:

- Measuring accuracy of the sampling frame and its impact on the survey frame;
- Measuring accuracy of data items;
- Evaluating accuracy of forecasting models or derivation procedures;
- Evaluating the accuracy of assumptions and limitations, calculations, and formulas used to create estimates; and,
- Evaluating accuracy of estimates by comparing estimates with other information sources.

Similar actions are applicable to other statistical data, secondary uses of nonstatistical data, and integrated data. However, when identifying and reporting quality for secondary-use data, over which agencies have less control, the needed information may not be available.

Accuracy and reliability for outputs of survey and other statistical data are frequently reported using traditional measures like standard errors and response rates that capture statistical bias and precision. Statistical bias and precision are described in Appendix B. Briefly, *statistical unbiasedness* refers to a condition where “the expected value of an estimate of a characteristic is equal to the true population value” (CNSTAT 2018, 40). The phrase “expected value of an estimate” refers to a value that would emerge if the population were sampled and re-sampled many times, the estimate calculated for each sample, and an average (expected value) of all sample estimates calculated from all the possible samples. *Statistical precision*, defined as the inverse of the variance, indicates the variability of an estimate and is related to the errors, or threats, that affect variance and standard errors, including sampling error and modeling error. Estimates with high statistical precision have lower variance and are more accurate than those with low statistical precision.

For integrated data products, incorporating errors for all inputs and processing steps is ideal but may not be feasible. Potential approaches for measuring accuracy and reliability for integrated data and similarly complex data products include: a) direct comparisons to an external gold standard; b) benchmarking with known subject-specific and other relevant information; and c) conducting sensitivity analyses, such as evaluations of critical processing, analysis decisions, and modeling assumptions.



Threats to Accuracy and Reliability

1. **Sampling error.** Sample surveys rely on a subset of a population to support estimates that represent the whole population. For a particular sample, estimates will differ from the true population characteristics, even if on average, over all possible samples, they are correct. These sample differences are caused by sampling error. All else being equal, a larger sample will result

**Accuracy
and
reliability**

in a smaller sampling error. Sampling error applies to surveys and integrated data products that include surveys. Sampling error affects the dimension granularity in the utility domain through its connection with sample size.

2. Nonresponse error and missing data. Nonresponse occurs when data are sought from but not provided by a unit. For surveys, this occurs when selected survey participants do not respond. Broken equipment creates missing data for sensors and other automated instruments. If units with particular characteristics are less likely to respond or less likely to have complete data, then estimates based on the collected data may be biased. A recent FCSM report describes methods used to assess the effects of nonresponse in federal surveys (FCSM 2020). Even when nonresponse and missing data are random, resulting estimates are based on a smaller number of collected observations.
3. Coverage error. For sample surveys, coverage error occurs when the sampling frame differs from the target population. Substantial coverage errors affect the utility of the data for inferences about the target population. Coverage error applies to secondary use of nonstatistical data when the universe from which the data were originally collected does not match the target population of the intended data product.
4. Measurement error. Measurement error is defined as the difference between the observed value of a variable and the true, but unobserved, value of that variable and as the variance (or standard deviation) of that difference.
5. Linkage error. Linkage error includes the possible types of errors that emerge from the linkage process itself, such as false matches or missed matches, that are not attributable to either data source independently.
6. Harmonization error. Harmonization errors arise when data elements have been defined, collected, or processed differently. For single source data products, harmonization errors affect consistency of the data over time. For integrated data products, differences in data elements across input sources can lead to harmonization error.
7. Modeling error. For modeled data products and outputs, modeling errors arise from inaccuracies in statistical model assumptions and from the effects of modeling decisions related to missing data, calibration variables and other constants, influential observations, and other factors. Modeling error can differ among statistical outputs from the same data product, including estimates produced for population domains and/or produced at various geographic and temporal resolutions (see additional threats to accuracy and reliability for geographic data below). In some instances, modeling errors can be evaluated and mitigated through combining models, a process sometimes called ensemble modeling, if the models are based on different but documented assumptions (*e.g.*, Centers for Disease Control and Prevention 2020, Seni and Elder 2010). Threats from modeling error can be exacerbated by measures to mitigate threats to timeliness and punctuality in the utility domain.
8. Processing error. Processing error occurs during the processes that convert collected data into data products, including data files, integrated data and data outputs. Commonly cited types of processing error include data entry, coding, editing, imputation, and analysis errors.

Accuracy and reliability

Threats from processing error can be exacerbated by measures to mitigate threats to timeliness and punctuality in the utility domain.

9. Additional Threats to Accuracy and Reliability Involving Geographic Data. Data that characterize geographic locations are affected by the geometry of boundaries, the ease of overextending data precision with geographic information systems technology, and the frequent use of neighborhood characteristics as a proxy for characteristics of individuals (Schmitt 1978). Threats to accuracy involving geographic data in the objectivity domain are closely tied to the dimension granularity in the utility domain.

Coherence is the ability of the statistical data to maintain common definitions, classification, and methodological processes; to align with external statistical standards; and to maintain consistency and comparability with other relevant data over time, across key domains, and when data originate from different internal and external sources. For surveys and statistical data collected by an agency, the use of validated questions and question-response research can increase coherence. For example, see the Q-bank collection of question evaluation studies (*e.g.*, NCHS 2020b). For nonstatistical data, toolkits are being developed that can be used as a starting point when assessing the acceptability of the coherence and other quality dimensions for data prior to use (Iwig 2013, Murphy and Konny 2017, Seeskin 2019).

Threats to coherence are summarized below. Each of these threats also affect accuracy and reliability and the dimensions accessibility and relevance in the utility domain.



Threats to Coherence

Coherence

1. Multiple sources of data and definitions. Information for the same constructs may be collected differently among surveys and other statistical data collections. For nonstatistical data acquired for secondary use, data collection methods can differ for similar constructs depending on the original purpose of the data, affecting coherence of the resulting data for a particular use.
2. Changes in data over time. Data sources, survey questions or collection instruments, and definitions can change over time within a data program. For data acquired for secondary use, changes in sources and content can be outside the control of the agency.
3. Changes in statistical and processing methods. Advances in statistical methodology and implementation may improve data products and outputs but can reduce coherence with prior releases of the same products and with other products using original methods.
4. Misalignment. Data collected for one purpose may not be coherent when used for another purpose.

Integrity

The National Research Council of the National Academies of Sciences, Engineering, and Medicine (2017) identified four principles for federal statistical agencies in its publication, “Principles and Practices for a Federal Statistical Agency.” These principles are fundamental to ensuring the integrity of official statistics: (1) objectivity of information produced, (2) credibility with those who use the data and information, (3) trust of those who provide information, and (5) independence within the government. Threats to integrity are associated with failing to uphold one or more of these principles. This section discusses threats to quality dimensions within the integrity domain: scientific integrity, credibility, computer and physical security, and confidentiality.

Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence. The “Statement of Commitment to Scientific Integrity by Principal Statistical Agencies” (Principal Statistical Agencies 2012) supports the breadth of factors that contribute to the integrity of official data products. Threats to scientific integrity vary in scope and impacts from political interference and obsolescence to computer-generated data.

To maintain scientific rigor, it is a continuing challenge to identify the most relevant newly available data sources and to develop methods that take full advantage of the additional information they provide. At the same time, consideration must be given to the potential for these data sources to be disrupted or for their content or scope to change, especially if they are obtained from the private sector. Changes and inconsistencies with external data sources could result in the scientific integrity of the released information being compromised if those changes are not considered in the production of the data product; unstable external source data also threatens coherence in the utility domain.

Scientific integrity



Threats to Scientific Integrity

1. **Political interference.** Interference in the publication of statistics, or even the appearance of such interference, can threaten the scientific integrity of a data product. Political interference is also a threat to credibility, another dimension in the integrity domain, and it affects perceptions of the data by users, a threat to relevance in the utility domain.
2. **Obsolescence.** The best available scientific and statistical methods must be used to obtain data and produce data products. If these methods do not evolve, over time processes that once led to scientifically rigorous data and information become dated and diminish the scientific integrity of the data products. Obsolescence also affects the dimension of credibility within the integrity domain and the dimensions of accuracy and reliability within the objectivity domain. Employing new methods, however, can reduce coherence with prior data products, another dimension in the objectivity domain.
3. **Computer-Generated Data.** Bots, software applications that run automated tasks or scripts over the internet, have the potential to provide inaccurate information to surveys (Dupuis *et al.* 2019, Chandler *et al.* 2017), possibly harming the objectivity and credibility of the published statistics. Bot generated data have been found among responses to crowd sourcing and opt-in

Scientific integrity

surveys for which respondents are being paid. The use of bots is likely to be low or nonexistent for probability sampled federal surveys but has yet to be fully studied for these and many other types of statistical and nonstatistical data. This threat can affect the credibility dimension in the utility domain, and the accuracy and reliability dimension in the objectivity domain.

Credibility is primarily a matter of trust. Do users trust the producer of the statistical products to provide an accurate and objective measure? As data produced by a trusted entity have more utility than the same data produced by a nontrusted entity, the organization's credibility transfers to the product. Credibility is derived from many aspects of quality inherent in a provider's range of data outputs, so the threats to credibility echo those of the other quality dimensions. Agencies can increase credibility by ensuring that the methods used are transparent, understandable, and rooted in accepted theory. Threats to credibility are varied and include the release of inaccurate and or unreliable products, competing data sources and methods, political interference, and obsolescence such as production of outdated products or use of outdated methods.



Some key practices for maintaining credibility have been recommended and include:

- Regular use of sound statistical methods (OMB 2014, 71615);
- Regularly evaluating of an agency's statistical products (OMB, 2014, 71615);
- Regularly providing transparency through clear descriptions of how data are collected or estimated, of assumptions that are made, and of any known data errors and limitations (National Research Council 2017, OMB 2014, 71615);
- Regularly making data and the information needed for users to work with the data widely available on an equal basis to all users (National Research Council 2017);
- Seeking input from all types of data users (National Research Council 2017);
- Making a strong commitment to professional practice (National Research Council 2017);
- Cultivating a reputation for good management and efficiency (European Commission 2011); and
- Demonstrating independence from undue political interference in the production, dissemination, and analysis of statistical data (European Commission 201, National Research Council 2017, OMB 2014).

Credibility



Threats to Credibility

1. Dissemination of inaccurate data products. Inaccuracies and errors impact the confidence that data users place in the data product and can affect their trust in other data products released by the agency. Users assign greater credibility to polls and other data products that have a history of accuracy (see, *e.g.*, Kennedy *et al.* 2018). All threats to accuracy and reliability, a dimension in the objectivity domain, affect the credibility of an agency and its data products.
2. Competing data sources and methods. Competing data sources and methods can diminish a data producer's credibility if they provide different answers to users' questions, particularly if

Credibility

differences are not understood. Competing data sources and methods is also a threat to relevance, another dimension in the utility domain. Documentation and communication about appropriate use of the data can mitigate this threat.

3. Competing data sources and methods. Competing data sources and methods can diminish a data producer's credibility if they provide different answers to users' questions, particularly if differences are not understood. Competing data sources and methods is also a threat to relevance, another dimension in the utility domain. Documentation and communication about appropriate use of the data can mitigate this threat.
4. Political interference. Interference in the publication of statistics, or even the appearance of such interference, can have profound repercussions for the credibility of a data provider and its actions. For example, when the technical report of the Advisory Commission to Study the Consumer Price Index (the "Boskin Commission") identified a potential upward bias in the Consumer Price Index in 1995, it included estimated impacts of this bias on the federal budget deficit. These implications were seized upon by the press and by political interest groups as suggesting possible political motives for the Commission's technical analysis (see, *e.g.*, Gordon 2000), reducing the credibility of both the Commission and the Consumer Price Index. Political interference is also a threat to scientific integrity, a dimension of the integrity domain, and it relates to the perceptions of users, a threat to relevance in the utility domain.
5. Obsolescence. For agencies to remain credible with those who use the data and information, the best available scientific and statistical methods must be used to obtain data and produce data products. If these methods do not evolve, over time processes that once led to credible data and information become dated and diminish the credibility of the agency and its data products. Obsolescence also affects the dimension of scientific integrity within the integrity domain and the dimensions of accuracy and reliability within the objectivity domain. Employing new methods, however, can reduce coherence with prior data products, another dimension in the objectivity domain.

Computer and physical security

Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification. This definition is based on prior OMB guidance (OMB 2006, 2002) and the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000). OMB guidance identifies several related best practices for survey data that can be applied to nonsurvey and nonstatistical data (OMB 2006):

- Establishing procedures and mechanisms to protect confidential information throughout the production and development process (Guideline 3.4.1, 19)
- Ensuring that data systems and electronic products are protected from unauthorized access storage systems that include confidential information, which are protected from unauthorized access (Guideline 3.4.2.2, 19)

Computer and physical security

- Ensuring that data files, network segments, servers, and desktop PCs are electronically secure from malicious software and intrusion (Guideline 3.4.2.3, 19) and
- Ensuring that access to electronic datasets with confidential information is provided on a need only basis controlled by the project manager (Guideline 2.4.3, 19)

Threats to computer and physical security are constantly growing and evolving and include those that originate outside of an agency, such as supply chain risks and other external threats, as well as insider threats and human error. The National Institute of Standards and Technology (NIST) maintains the Federal Information Security Management Act (FISMA) Implementation Project² which provides the overarching framework all federal agencies need to follow. Some primary references are NIST SP 800-37r2³ and NIST SP 800-53r5⁴ that provide additional detail critical for this topic area.



Threats to Computer and Physical Security

1. Supply chain risk. The supply chain can be compromised by several threat vectors. Bad actors may try to sneak malicious devices on the hardware during production, modified post production while in transit, at the data center, or while in use. To follow U.S. laws and regulations, some equipment cannot be obtained when manufactured in certain countries or points of origin. The 2020 Covid-19 pandemic also exposed the weakness in availability of parts that can create service disruptions/outages when parts need replacement.
2. Human error. Human errors are those that are accidental disclosures or disruptions that are not intentional or willful, malicious acts. Malicious acts may depend on human error such as clicking on links or opening email attachments. Disclosure due to accidental loss of physical devices (*i.e.* smart phone, laptop, hardware tokens, etc.) can lead to attack. Other errors include not following policy, training, or other required rules that require compliance. For IT systems, misconfiguration or failure to apply required protections to IT assets can lead to data breaches and other disruptive attacks.
3. Insider Threat. Attempts by employees or trusted individuals to gain unauthorized access to systems and data are insider threats. Acts include, but are not limited to, exfiltration or theft of sensitive data and/or equipment, destruction of property, denial of service, and installation of unauthorized software that performs malicious or illegal acts.
4. External Threats. External actors seeking to gain illegal access, disrupt, or destroy IT systems and data through a variety of threats and tactics are external threats. Common examples of external threats include phishing, malicious hacking, denial of service attacks, logic bombs, and ransomware. These threats could include staking out physical location, cataloging operations and vulnerability, and attempts to gain access to the facility to either gain access or test detection capabilities.

2. Federal Information Security Management Act (FISMA) Implementation Project <https://www.nist.gov/programs-projects/federal-information-security-management-act-fisma-implementation-project>

3. Risk Management Framework for Information Systems and Organizations Revision 2. <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-37r2.pdf>

4. Security and Privacy Controls for Information Systems and Organizations. <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5-draft.pdf>

Confidentiality

Confidentiality refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party. This definition is from the Evidence Act (Pub. L. No. 115-435, 132 Stat. 5529, 2018, codified at 44 USC 3563(d)(4)) and is consistent with requirements that individually-identifiable data be protected from unauthorized disclosure in order to uphold its confidentiality (OMB 2006). Individually-identifiable data are data that permit the identity of the respondent or entity to which the information apply to be reasonably inferred by either direct or indirect means. A growing literature has acknowledged and formalized the fact that each data product published from a given dataset will incur some expenditure of the underlying dataset's "privacy budget" (see, *e.g.*, Abowd and Schmutte 2019).

Threats to confidentiality increase with the level of detail provided by the data, including the granularity of the data and the number of data elements included in microdata files. Rare or unusual population characteristics or phenomena included in a data product can also increase threats to confidentiality. Methods for avoiding disclosure of confidential data and the interaction of those methods with granularity and other data quality elements are described under utility. Protection of confidentiality is increasingly challenged by big data analytics that include increasingly sophisticated data science techniques that can reverse engineer individual records from related datasets. FCSM is continually updating its guide to best practice in protecting confidentiality, adopting a more flexible structure to permit more frequent updates (FCSM 2005, OMB 2019).



Threats to Confidentiality

1. Granularity. Disclosure risks increase with granularity. Data products, including data files and outputs that provide information for small population groups, small population geographic units, and narrow time intervals increase disclosure risks. This threat to confidentiality is mitigated by reducing the granularity with which data are released, which increases the threat to data quality in the utility domain through the granularity dimension.
2. Large number of data elements in microdata. Disclosure risks for microdata files increase with the number of available data elements and the resulting likelihood of unique or near-unique records. This threat is increased when the file or output includes rare or unlikely combinations defined by multiple elements and for linked data products which increase the data elements associated with individual units. To mitigate this threat data products undergo rigorous process to evaluate disclosure risk prior to release and various methods of disclosure reduction are applied to the files that are released. These mitigations can impact the utility of the files.

References

- Abowd JM, Schmutte IM. 2019. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*. 109(1):171-202.
- Adrian D. 2012. A Model-Based Approach to Forecasting Corn and Soybean Yields. *Proceedings of the Fourth International Conference on Establishment Surveys, 2012*. Alexandria VA: American Statistical Association. <http://www.amstat.org/meetings/ices/2012/papers/302190.pdf>.
- Beede D, Powers R. 2015. Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data. 10.13140/RG.2.1.3354.8888.
- Biemer PP, de Leeuw E, Eckman S, Edwards B, Kreuter F, Lyberg LL, Tucker C, West BT (eds). 2017. *Total Survey Error in Practice*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc. Hoboken, New Jersey.
- Blumenfeld J. 2016. NASA EOSDIS Role in the Big Earth Data Initiative, Earthdata. National Air and Space Administration, October 17, 2016. Available at: <https://earthdata.nasa.gov/learn/articles/tools-and-technology-articles/eos-dis-role-in-bedi>
- BTS. 2020. The Week in Transportation. Available at <https://www.bts.gov/newsroom/week-transportation>.
- Cajner T, Crane LD, Decker RA, Hamins-Puertolas A, Kurz C. 2019. Improving the Accuracy of Economic Measurement with Multiple Data Sources: The Case of Payroll Employment Data (No. w26033). National Bureau of Economic Research.
- U.S. Census Bureau. 2020. Federal Statistical Research Data Centers. Found at <https://www.census.gov/fsrdc> (accessed May 15, 2020).
- Centers for Disease Control and Prevention. 2020. Covid-19 forecasts. <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html>.
- Chandler JJ, Paolacci G. 2017. Lie for a Dime: When Most Prescreening Responses Are Honest but Most Study Participants Are Impostors. *Social Psychological and Personality Science* 8(5): 500508. <https://doi.org/10.1177/1948550617698203>.
- CNSTAT. National Academies of Sciences, Engineering, and Medicine. 2019a. Reproducibility and replicability in science. National Academies Press.
- CNSTAT. National Academies of Sciences, Engineering, and Medicine. 2019b. Methods to Foster Transparency and Reproducibility of Federal Statistics: Proceedings of a Workshop. Washington, DC: The National Academies Press. Available at: <https://www.nap.edu/catalog/25305/methods-to-foster-transparency-and-reproducibility-of-federal-statistics-proceedings>.
- Cruze NB. 2015. Integrating Survey Data with Auxiliary Sources of Information to Estimate Crop Yields. In *JSM Proceedings*, Survey Research Methods Section. Alexandria VA: American Statistical Association. 565-578. <https://www.amstat.org/sections/srms/proceedings/y2015/files/233920.pdf>.
- Cruze NB, Benecha HK. 2017. A model-based approach to crop yield forecasting. In *JSM Proceedings*, Survey Research Methods Section. Alexandria VA: American Statistical Association. 2724-2733.
- Czajka JL, Stange M. 2018. Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines." Washington, DC: Mathematica Policy Research, April 27, 2018.

- Dupuis M, Meier E, Cuneo F. 2019. Behaviour Research Methods 51:2228. <https://doi.org/10.3758/s13428-018-1103-y>.
- European Commission. 2011. European Statistics Code of Practice. Brussels, Belgium: European Statistical System Committee, September 28, 2011. Available at: <https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice>.
- Eurostat. 2020. European Statistical System Handbook for Quality and Metadata Reports, 2020 edition. 2020. Luxembourg: Publications Office of the European Union. Available at <https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf/bf98fd32-f17c-31e2-8c7f-ad41eca91783>.
- Evans, M, He, Y, Yevseyeva, I, and Janicke, H. 2018. Analysis of published public sector information security incidents and breaches to establish the proportions of human error. Proceedings of the 12th International Symposium on Human Aspects of Information Security & Assurance/ 191-202.
- FCSM. 2001. Measuring and reporting sources of error in surveys. Washington, DC: U.S. OMB (Statistical Policy Working Paper 31).
- FCSM. 2005. Report on Statistical Disclosure Limitation Methodology. Washington, DC: U.S. OMB (Statistical Policy Working Paper 22). Available at <https://nces.ed.gov/FCSM/pdf/spwp22.pdf>.
- FCSM. 2020. A Systematic Review of Nonresponse Bias Studies in Federally Sponsored Surveys. Washington, DC. https://nces.ed.gov/fcsm/pdf/A_Systematic_Review_of_Nonresponse_Bias_Studies_Federally_Sponsored_SurveysFCSM_20_02_032920.pdf.
- Glaeser EL, Kim H, Luca M. 2017. Nowcasting the local economy: Using yelp data to measure economic activity (No. w24010). National Bureau of Economic Research.
- Gordon RJ. 2000. The Boskin Commission report and its aftermath (No. w7759). National bureau of economic research.
- Groves R, Fowler F, Couper M, Lepkowski J, Singer E, Tourangeau R. 2009. Survey Methodology, 2nd ed. Hoboken, New Jersey: John Wiley & Sons.
- HHS Data Council. 2001. Improving Data for Decision-Making: HHS Data Collection Strategies for a Transformed Health System, December 21, 2011. Available at: <https://aspe.hhs.gov/basic-report/improving-data-decision-making-hhs-data-collection-strategies-transformed-health-system>
- ICSP. 2018. Principles for Modernizing Production of Federal Statistics. Available at <https://nces.ed.gov/fcsm/pdf/Principles.pdf>.
- Iwig W, Berning M, Marck P, Prell M. 2013. Data quality assessment tool for administrative data. Prepared for a subcommittee of the Federal Committee on Statistical Methodology, Washington, DC. <https://nces.ed.gov/FCSM/pdf/DataQualityAssessmentTool.pdf>
- Kennedy C, Blumenthal M, Clement S, Clinton JD, Durand C, Franklin C, McGeeney K *et al.* 2018. An evaluation of the 2016 election polls in the United States. Public Opinion Quarterly 82(1):1-33.
- Murphy B, Konny C. 2017. Quality Considerations for Administrative Data Used for the Producer Price Index (PPI) and Consumer Price Index (CPI) development. Presented to the FCSM/WSS Workshop on Quality of Integrated Data Reporting on Quality Issues: Input data, December 1, 2017. Available at: www.washstat.org/presentations.
- National Research Council. National Academies of Sciences, Engineering, and Medicine. 2017. Principles and Practices for a Federal Statistical Agency: Sixth Edition. Washington, DC: The National Academies Press. <https://doi.org/10.17226/24810>.

NCHS. 2020a. Excess Deaths Associated with COVID-19. Provisional Death Counts for Coronavirus Disease (COVID-19). Available at https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm

NCHS. Collaborative Center for Questionnaire Design and Evaluation Research. 2020b. Q-bank. Available at <https://wwwn.cdc.gov/qbank/Home.aspx>. Accessed May 1, 2020.

OMB. 2002. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies, 67 FR 8451.

OMB. 1985. Statistical Policy Directive No. 3: Compilation, Release, and Evaluation of Principal Federal Economic Indicators (50 FR 38933, September 25, 1985). <https://s3.amazonaws.com/archives.federalregister.gov/issue/slice/1985/9/25/38908-38934.pdf#page=25>

OMB. 2006. Standards and Guidelines for Statistical Surveys. September 2006. Available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

OMB. 2008. Statistical Policy Directive No. 4. Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies (73 FR 12621, March 7, 2008). Available at: <https://www.govinfo.gov/content/pkg/FR-2008-03-07/pdf/E8-4570.pdf>.

OMB. 2014. Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units (79 FR 71609, Dec 2, 2014).

OMB. 2019. “2020 Federal Data Strategy Action Plan”, available at <https://strategy.data.gov/action-plan/>.

Prell M, Chapman C, Adeshiyan S, Fixler D, Garin T, Mirel L, Phipps P. 2019. Transparent Reporting for Integrated Data Quality: Practices of Seven Federal Statistical Agencies. FCSM 19-01. September 2019. <https://nces.ed.gov/fcsm/>

Principal Statistical Agencies, 2012. Statement of Commitment to Scientific Integrity (last modified on May 23, 2012). Accessed on June 24, 2020 from: <https://www.bls.gov/bls/integrity.htm>

Pub. L. No. 106-554, § 515(a). 2000. Information Quality Act.

Pub. L. No. 115-435, 132 Stat. 5529. 2018. Foundations for Evidence-Based Policymaking Act of 2018.

Rabe M, Ortman J, Pascale J, Ikeda M, Cantwell P, Heimel S, Poehler E *et al.* 2017. MEMORANDUM FOR Victoria Velkoff Chief, American Community Survey Office From: David Waddington Chief, Social, Economic, and Housing Statistics Division (SEHSD) Prepared by Edward Berchick.

Schildkraut JL, Baker CA, Cho KN, Reuss KL. 2015. The National Compensation Survey and the Affordable Care Act: preserving quality health care data, Monthly Labor Review, U.S. Bureau of Labor Statistics, April 2015, <https://doi.org/10.21916/mlr.2015.9>.

Schmitt RR. 1978. Threats to Validity Involving Geographic Space, Journal of Socio-Economic Planning Sciences, 12: 191-195.

Statistical Community of Practice (SCOPE) Metadata team. 2020. Metadata Systems for the U.S. Statistical Agencies, in Plain Language. Found at https://nces.ed.gov/fcsm/pdf/Metadata_projects_plain_US_federal_statistics.pdf.

Seni G, Elder JF. 2010. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Morgan and Claypool.



4 Best Practices for Identifying and Reporting Data Quality

The Importance of Reporting Data Quality

The ICSP has stated that “agencies should work to adopt a common language and framework for reporting on the quality of data sets and derivative information they disseminate” (ICSP 2018). The framework described in this report is designed to fill that need. The framework establishes core concepts grounded in the prevailing literature and international experiences, which can be flexibly applied to accommodate agencies’ needs and priorities.

Data and analyses can provide powerful insight and understanding for decision makers, but they can just as easily misguide decision makers. To serve the rapidly evolving world of data-driven decision-making, data producers and analysts must identify and report the quality of the data they produce so that the data products can be used effectively.

No data are perfect. The increasing use of integrated data and nonstatistical data for statistical purposes sometimes compounds and sometimes ameliorates those imperfections. Ideal approaches for identifying and reporting data quality necessarily differ among data products and the complexity of these activities is higher for complicated data products, such as integrated data. For integrated data in particular, ICSP (2018) emphasizes the need to provide data users with contextual information about source data, the impact of disclosure limitation treatments, integration methods, when applicable, and the assumptions, defaults and uncertainties that underlie the methods used to convert source and integrated data into data products.

The framework encompasses a broad concept of data quality that spans many dimensions, allowing data producers and analysts, data users and other stakeholders to think holistically about the strengths and weaknesses of data products and assess trade-offs among the dimensions when making decisions. As shown in the discussion of data quality considerations in Chapter 3 and described in Appendix B (Accuracy and Reliability of Integrated Data), the framework can be used to identify threats and manage trade-offs among them.

The framework was developed for statistical data, data designed and collected for statistical purposes (*i.e.* surveys and censuses), and nonstatistical data. Nonstatistical data, data originally collected for administrative or other purposes, are increasingly acquired for secondary use and include administrative records, data from satellites, sensors and other monitors, and web scraping. The framework can also be used for integrated data, which are produced using statistical data, nonstatistical data or both.

The framework can be used to assess data quality threats for all types of data products. These data products include data collections and data files as well as the statistical information produced from them, such as tables, graphics and estimates. The framework can be used to evaluate data produced through integration, modeling, harmonization and other analyses where identifying threats to each dimension of data quality in the framework can be done for all steps of the production process. Integrated data products, for example, include the quality threats for each input source and those resulting from integration methods (Zhang 2012). Additional quality threats can result from the analytic methods used to produce data outputs.

Reporting data quality is necessarily nested, from comprehensive internal documentation for current and future data stewards and program managers to the most relevant extracts of that documentation for different external uses and users. Comprehensive and accessible documentation maintained by data producers is the foundation for the data quality information provided for users. This chapter briefly summarizes best practices developed by the FCSM for how the framework can be used to identify and report data quality. Best practices were developed following from the factors that affect data quality described earlier in this report, building on earlier FCSM reports focusing on the accuracy of survey data (*e.g.* FCSM 2001), and extending these to all data as called for by the Federal Data Strategy (OMB 2019a, 2019b). The implementation of best practices includes assessment and documentation of the primary threats to data quality in each of its dimensions, including effects of mitigations and trade-offs among them.



Identifying Threats to Data Quality

The framework can be used to regularly and systematically consider potential threats to the utility, objectivity and integrity of the data, and their associated dimensions. In addition, uncertainties about the ability to anticipate threats, as well as actions taken to detect and address threats, can be assessed and identified. Even though some identified threats may not warrant special scrutiny or inclusion in a data product's documentation, assessing all dimensions is still important. Furthermore, using the framework to systematically identify threats to data quality facilitates the management of trade-offs among them and their mitigations for decision making.



Best practices for identifying threats to data quality include:


- Consider threats to data quality when developing a new data collection or data product, including threats from new sources and methods. Regularly identify threats to data quality for ongoing data collections and methods. Identification of threats and management of trade-offs among them should be done in the context of the purpose of the data and all other identified threats.
- For integrated data products, identify threats to data quality for all source data in the context of the purpose of the integrated data. For a particular data source, the most salient threats for use in an integrated data product may differ from those for its original purpose.
- For integrated data products, identify threats to quality resulting from the integration method, including record linkage, modeling, and harmonization, using appropriate statistical methods. Management of trade-offs among identified threats and measures to mitigate their impact will depend on the primary uses of the integrated data product.
- For data outputs, including estimates in tables and reports, identify threats to data quality that follow from the quality of the data collection as well as additional threats resulting from statistical methods used to produce the outputs. For data outputs, management of trade-offs among threats and mitigation measures should be considered in the context of the purpose of the data and all identified threats. For a specific data output and purpose, the most salient data quality dimensions can differ from those identified for the data collection and may differ from those from other outputs from the data collection.

Reporting Data Quality

The ideal level of data quality reporting depends on the users of the information and the data product being documented. The framework can be used to describe identified threats to data quality, manage trade-offs among identified threats, and employ any actions needed to mitigate the impact of identified threats. Quality reporting can take many forms, from detailed documentation about data collections to technical notes or footnotes accompanying a published statistic. Reporting varies, with some data programs following detailed guidance on structure and content for documentation (e.g., OMB 2006, Seastrom 2012). Reporting for a particular data output will build from the detailed documentation for the data collection from which it is derived and will include the documentation of additional quality assessments for any integration methods and statistical methods used in its production.




Best practices for reporting data quality include:

 <p>1</p>	<p>Create detailed technical documentation about all aspects of the data program product to meet the programmatic needs for product continuity and dissemination. Documentation should describe the main features of the data product, including its purpose, principal uses, population and sub-population coverage, the time frames for which it is available, and the expected periodicity of its production. It should include technical operations and production details, methods used in processing, including imputations and weighting, editing, integration methods and evaluations, and data dictionaries. In addition, any changes made to data files to ensure confidentiality and any impacts on inference should be described. For ongoing data products, changes to key constructs, methods, outputs, and their impacts on data quality should be described.</p>
 <p>2</p>	<p>Summarize the quality of the data product in its internal documentation, including identified threats, management of trade-offs among them, and any mitigation measures employed. Although not all dimensions of the framework will apply to each data product and detailed quality assessments may only be needed periodically for ongoing data products, the results of regular assessments of applicable dimensions of the framework should be documented. By reporting data quality for a broad set of dimensions, the internal documentation can be used to tailor additional user reports for users on data products and outputs with the most relevant and valued data quality information.</p>
 <p>3</p>	<p>Consider the quality of each input data source and all processing steps when documenting integrated data. Documentation should include the methods used for additional processing of source data needed for integration, including imputations and editing, if needed, and relevant methodological information about integration methods. It should include detailed evaluations of identified threats to the integrated quality of the data, particularly threats most relevant for the integrated data. Because the intended use of an integrated data product may differ from the original purposes associated with the input data, data quality for each input source should be documented for its use in the integrated data product.</p>


 <p>4</p>	<p>Provide to data users an overview of the data product and its purpose, its principal uses, its population and sub-population coverage, the time frames for which it is available, and the expected periodicity of its production. For some data outputs, a Frequently Asked Questions section may be used to paraphrase such information. Documentation for data users should include the most relevant data quality issues, including the strengths and weaknesses of the data, any mitigation approaches employed to address quality threats, and implications for the use of the data and its audience. Because the purpose of a particular data product or output may not have been considered in the original quality evaluation of a data collection, any additional (or reduced) threats to data quality identified its intended use and audience should be documented. Threats resulting from statistical analysis should be evaluated and reported. Measures of uncertainty such as standard errors and intervals should always be available for estimates in tables, reports, and other disseminations.</p>
 <p>5</p>	<p>Report data quality to users with various levels of detail. Although detailed technical documentation is needed within a data collection program for planning, decision making, and continuity, extracts of the most salient qualities of the technical documentation can be created for various data products and audiences. The level of detail needed for documentation provided to “power-users” of microdata files and statistical products differs from that needed for occasional users of tables and reports. The detail needed for complex integrated data sets may differ from the detail needed for small surveys. In all cases, the availability of additional, relevant detailed documentation should be identified.</p>
 <p>6</p>	<p>Prepare a high-level summary encapsulating the key quality issues relevant to the data product and its primary users. In a few sentences, the summary would provide an overview of the data product’s origin and describe its suitability for a particular use or uses and how likely it is that key data outputs would lead to misleading information.</p>

Moving Forward

As articulated in the introduction of this report, effective understanding of data quality is essential to data-driven decisions by public officials, private business, and the public. This report summarizes the state of practice in identifying and reporting data quality issues consistent with a framework established in the Information Quality Act (Pub. L. No. 106-554, § 515(a), 2000) and described best practices put forth by FCSM for identifying and reporting quality threats. In addition, several approaches and methods have been identified that have yet to be adopted within most, if not all, federal statistical agencies, or that require additional research before they can be used extensively, including:

 <p>1</p>	<p>Methods for evaluating the quality of official statistics based on integrated data. For official statistics derived directly from surveys, metrics for assessing data quality based on the sources of error are well established (<i>e.g.</i> FCSM 2001). For official statistics produced using integrated data, the quality depends on the quality of the primary sources and how they are combined. For official statistics developed from integrated data, new measures of data quality are needed (Agafitei <i>et al.</i> 2015, Keller <i>et al.</i> 2017). Metrics capturing the added value, if any, from integrated data compared to survey data could be insightful and guide future efforts;</p>
 <p>2</p>	<p>Identifying determinants of response and measurement error in administrative data and other nonsurvey data. Researchers have developed a deep knowledge base about how and why survey respondents provide accurate responses, enabling surveys to be designed to elicit the best information (Groves <i>et al.</i> 2009). As agencies rely more and more on administrative data and other nonsurvey data to create statistical outputs, a similar set of research applied to these nonsurvey collections is likely to be valuable. As an example, North American Industry Classification System (NAICS) codes comprise a business classification based on industry production processes. NAICS codes are self-reported on tax forms and due to their complexity may be subject to considerable error⁵. The Internal Revenue Service Statistics of Income (SOI) uses an intensive manual validation process for their published statistical data. Statistical models developed using SOI data can be used to improve NAICS codes on individual, corporate and partnership returns that are missing or incorrect;</p>
 <p>3</p>	<p>Methods for assessing coverage errors when data are not sampled according to a pre-determined design. The challenges of measuring the coverage of sampling frames and adjusting for undercoverage (or overcoverage) have been well studied (Groves, 1989). When adopting methods for integrated data using administrative data, the data may represent a nonprobability sample; other data may be intended to be complete for a specific population, but are not. New or enhanced approaches to measuring and adjusting for coverage that can be adopted in both research and production environments would be valuable. As an example, linking data to a disease registry may lack coverage if certain states are not represented in the registry;</p>

5. The estimated error rates for administrative data for 2007–2016 were approximately 20 percent for Forms 1040 Schedule C (Profit or Loss from Business).

 <p>4</p>	<p>Methods for assessing the interaction between different types of errors. For example, how do missingness and linkage errors interact? How do harmonization errors and coverage errors interact? What are the tradeoffs between timeliness and model error assessment? Sensitivity analyses are a valuable tool in this process. Yet, at the present time, these analyses generally evaluate the effect of one error at a time and a representative truth source to be used as comparison is rarely available (Harron et al. 2017). Developing designs and subsequent analyses for sensitivity studies that allow these more complex interactions to be evaluated is important (Saltelli et al. 2006). Increasing use of ensemble methods in data analytics can introduce hidden errors as outputs from initial models, which are subject to estimation error, are integrated into blended data and used as inputs into other analytical methods (Singh, et al. 2020);</p>
 <p>5</p>	<p>For integrated data, methods to quantify uncertainty of the outputs given the errors across all input sources, including errors from the integration steps and the final analyses leading to the outputs. The total survey error paradigm provides a framework for quantifying the uncertainty of survey-based estimators. When reporting uncertainty of these statistics, the measure of error accounts for sampling error and some nonsurvey errors such as nonresponse and undercoverage but not for other nonsampling errors such as linkage error (Groves and Lyberg 2010). The challenge of accurately reflecting uncertainty is even greater for integrated data and may require new methods (Pferffermann 2015). Developing guidelines and/or methods that allow an agency to determine the best approach for a particular application would be helpful;</p>
 <p>6</p>	<p>Methods for disclosure risk protection and documentation of disclosure risk protection. A variety of statistical disclosure methods have been used within agencies, including data swapping, coarsening, noise infusion, and synthetic data. With the increase use of integrated data, these are being re-evaluated. Differential privacy, which focuses on quantifying the risk of disclosure, utilizes a metric for controlling that risk. The Census Bureau is developing algorithms for the 2020 Census that will allow them to adopt a differential privacy approach (Abowd 2018, Dwork 2019). Other approaches to statistical disclosure limitation, such as online table builders, are also being developed (Shlomo et al. 2019). There is a growing recognition of the importance of controlling disclosure risks using differential disclosure methods, especially as more data become publicly available. Many agencies would benefit by having approaches identified for assessing what method should be adopted and for adapting the selected method to a particular application;</p>
 <p>7</p>	<p>Best practices for communicating quality across various audiences with new sources of data. A major trend in official statistics is the increased use of graphics and interactive dissemination tools. Sometimes the quality of the statistics become less evident in the process. However, research is providing insights into how to communicate uncertainty graphically (Potter et al. 2012). Guidelines for best practices would be useful; and</p>



8

New templates and related tools for recording internal data quality documentation and converting that documentation into data quality reports and data quality components of standard metadata that take advantage of new technologies. As an example, using administrative data for analytics often requires an understanding of the business processes behind the data (Singh, et al. 2020). Some appropriate metrics for assessing the quality dimension of accuracy and reliability within the objectivity domain are well established. Identifying the best metrics for all of the dimensions is a crucial step toward being able to establish templates and tools that can easily lead to data quality reports and data quality components of standard metadata when analyses are based on integrated data.

The history of work to evaluate data quality includes many successes and a few ideas (such as data quality profiles) that have not completely fulfilled their promise. Future evolution will be rapid and positive if data producers embrace a culture of continuous improvement and channel their experiences, both successful and less successful, into a learning agenda as recommended by the Commission of Evidence-Based Policymaking (2017) and reinforced by the Foundations for Evidence-Based Policymaking Act of 2018 (Pub. L. No. 115-435, 132 Stat. 5529, 2018). Although statistics are a key part of learning from experiments, experiments by data producers and analysts in dealing with data quality are integral to learning how to identify, measure and report data quality more effectively. State-of-the-art methods for measuring and reporting data quality will advance if data producers and analysts cultivate opportunities to collaborate and share their experiences to advance our efforts to promote data quality and transparency in reporting for federal data.

References

- Abowd J. 2018. The U.S. Census Bureau adopts differential privacy. KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. <https://dl.acm.org/doi/abs/10.1145/3219819.3226070>.
- Agafitei M, Gras F, Kloek W, Reis F, Văju S. 2015. Measuring output quality for multisource statistics in official statistics: Some directions. *Statistical Journal of the International Association of Official Statistics*, Volume 31, Pages 203-211. <https://content.iospress.com/download/statistical-journal-of-the-iaos/sji902?id=statistical-journal-of-the-iaos/sji902>.
- Commission on Evidence-Based Policymaking. 2017. The Promise of Evidence-Based Policymaking, available at <https://www.cep.gov/cep-final-report.html>.
- Dwork C. 2019. Differential privacy and the U.S. Census. PODS '19: Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. <https://dl.acm.org/doi/abs/10.1145/3294052.3322188>
- FCSM. 2001. Measuring and reporting sources of error in surveys. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 31).
- Groves R. 1989. *Survey Errors and Survey Costs*. Hoboken, New Jersey: John Wiley & Sons.
- Groves R, Fowler F, Couper M, Lepkowski J, Singer E, Tourangeau R. 2009. *Survey Methodology*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons.
- Groves R, Lyberg L. 2010. Total survey error: Past, present and future. *Public Opinion Quarterly*, Volume 74, Issue 5, Pages 849-879.
- Harron, KL, Doidge, JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, van der Meulen JH. 2017. A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology*, Volume 46, Issue 5, October 2017, Pages 1699–1710, <https://doi.org/10.1093/ije/dyx177>.
- ICSP. 2018. Principles for Modernizing Production of Federal Statistics, 2018. Available at <https://nces.ed.gov/fcsm/pdf/Principles.pdf>.
- Keller S, Korkmaz G, Orr M, Schroeder A, Shipp S. 2017. The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application*, Volume 4, Issue 1, Pages 85-108. file:///C:/Users/LJYou/Downloads/deps_191048.pdf.
- OMB. 2006. Standards and Guidelines for Statistical Surveys. September 2006. Available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.
- OMB. 2019a. Memorandum M-19-18, “Federal Data Strategy - A Framework for Consistency,” available at <https://www.whitehouse.gov/wp-content/uploads/2019/06/M-19-18.pdf>.
- OMB. 2019b. 2020 Federal Data Strategy Action Plan, available at <https://strategy.data.gov/action-plan/>.
- Pfeffermann D. 2015. Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, Volume 3, Issue 4, Pages 425-483.
- Potter K, Rosen P, and Johnson CR. 2012. From quantification to visualization: A taxonomy of uncertainty visualization approaches. A Dienstfrey and RF Boisvert (Eds): *WoCoUQ 2011, IFIP AICT 377*, Pages 226-249, available at https://link.springer.com/content/pdf/10.1007/978-3-642-32677-6_15.pdf.

Pub. L. No. 106-554, § 515(a). 2000. Information Quality Act.

Pub. L. No. 115-435, 132 Stat. 5529. 2018. Foundations for Evidence-Based Policymaking Act of 2018.

Saltelli A, Ratto M, Tarantola S, Campolongo F. 2006. Sensitivity analysis practices: Strategies for model-based inference. *Reliability Engineering and System Safety*, Volume 91, Pages 1109-1125

Seastrom M, 2012. NCES Statistical Standards, Standard 7-2 Survey Documentation in Reports, NCES 2014097.

Shlomo N, Krenzke T, Li J. 2019. Comparison of three post-tabular confidentiality approaches for survey weighted frequency tables. *Transactions on Data Privacy*, Volume 12, Pages 145-168.

Singh L, Traugott M, Bode L, Budak C, Davis-Kean PE *et al.* 2020. Data Blending: Haven't We Been Doing This for Years? White paper from the Massive Data Initiative. Georgetown University. https://www.jonathanmladd.com/uploads/5/3/6/6/5366295/mdi_data_blending_white_paper_-_april2020.pdf.

Zhang, LC. 2012. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41-63.

Appendix A. Additional Background on Data Quality

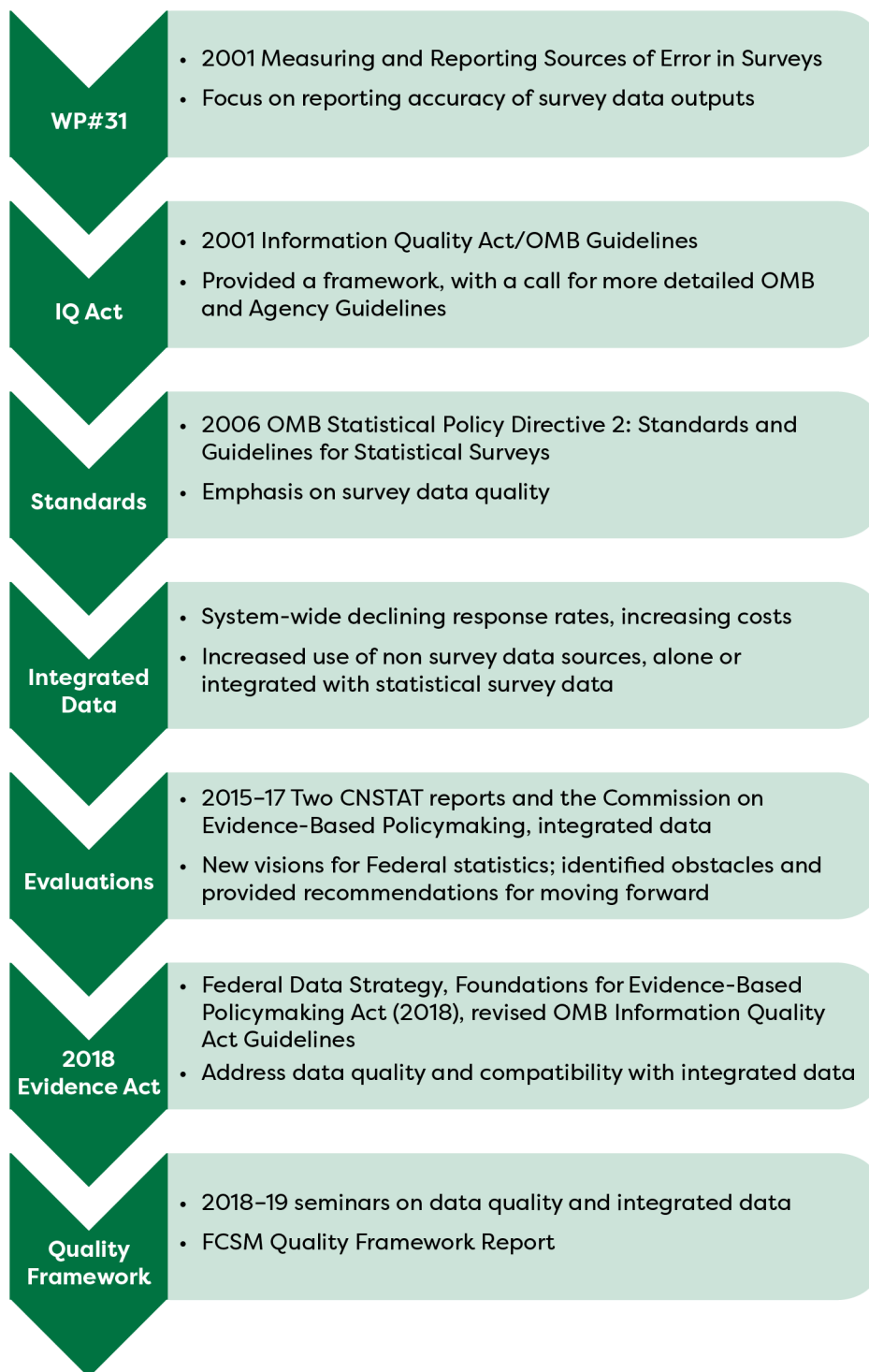
A.1 Summary

This Appendix is intended to link several strands of previous work relating to the responsibilities of federal agencies to address data quality. This appendix identifies key touchpoints from this previous work, providing high level summaries of the related documents. It proceeds, chronologically, establishing a selected history of the topic's development over the last 20 years. A summary timeline is shown in Figure A below.

A.2 A Foundation for Assessing Data Quality in Surveys (1978-2001)

As it is a fundamental concern to the Federal Statistical System (FSS), data quality has been the subject of many previous FCSM technical reports. Such reports have provided insights on data quality and error measurement in employment data from the Current Population Survey (FCSM 1978), establishment surveys (FCSM, 1988), survey coverage (FCSM, 1990a), data editing (FCSM 1990b), and reporting error in analytic publications (Atkinson *et al.* 1999). In 2001, FCSM issued Working Paper 31, "Measuring and Reporting Sources of Errors in Surveys," which provided a general discussion of sources of error in data collection programs (FCSM 2001). The 2001 report drew upon the lessons from previous FCSM reports as well as what was, by then, a well-established understanding of the impact of various types of survey errors on the accuracy of survey-based estimates. Working Paper 31 started with a discussion of the data quality policies and guidelines in place in 2001 and then focused on five error types: sampling error, nonresponse error, coverage error, measurement error, and processing error. For each of these error types, it briefly discussed the methods used to measure the error source, reviewed how federal statistical agencies reported on errors, and made recommendations for federal statistical agencies' future reporting in two types of outputs: analytic reports and technical reports. The paper with a discussion of measurement and reporting of Total Survey Error. Working Paper 31 noted that data quality is a multidimensional concept that includes accuracy, relevance, timeliness and accessibility, but focused on the accuracy dimension (FCSM 2001).

Figure A. Data Quality Milestones 2001–2020



A.3 Statutory and Policy Requirements (2001-2006)

The Information Quality Act, issued as Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 (Pub. L. No. 106-554, § 515(a), 2000) directed OMB to issue government-wide guidelines that “provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, Objectivity, Utility, and Integrity of information (including statistical information) disseminated by Federal agencies.” Acknowledging the wide range of potential applications of these principles to specific circumstances, the Act stipulated that OMB guidance should require each applicable agency to issue its own implementation guidelines. OMB issued guidelines in 2002 to implement the Act (OMB 2002a, 2002b).

In the OMB guidance, information quality is described as an overarching concept that includes Objectivity, Utility, and Integrity of information. Objectivity refers to presenting accurate, reliable, and unbiased information in an accurate, clear, complete and unbiased manner. Utility refers to the usefulness of the information to the intended users. Integrity refers to providing the data security required to protect information from unauthorized access or revision through corruption or falsification (OMB 2002a).

During 2002, OMB issued several additional policies that revised, clarified, and expanded its initial guidance. The June 4, 2002 OMB guidance, (OMB 2002b) provided cites to each statistical organization’s draft guidelines and described some basic features of how the Nation’s principal statistical organizations would be responsive to the OMB guidelines. These guidelines were intended to be included as part of the responses from the Departments and agencies in which the statistical organizations are located. Departments and agencies were required to issue their own implementing guidelines. Features included a commitment to quality and professional standards of practice, such as:

- Using modern statistical theory and practice in all technical work;
- Developing strong staff expertise in the disciplines relevant to its mission;
- Implementing ongoing quality assurance programs to improve data validity and reliability and to improve the processes of compiling, editing, and analyzing data; and
- Developing a strong and continuing relationship with appropriate professional organizations in the fields of statistics and relevant subject-matter areas.

The guidance called for applying high standards of performance to the following activities:

- Development of concepts and methods;
- Planning and design of surveys and other means of collecting data;
- Collection of data;
- Processing and editing of data;
- Analysis of data;
- Production of estimates or projections;
- Establishment of review procedures; and
- Dissemination of data by published reports, electronic files and other media requested by users.

An FCSM Committee used this list of activities as a backbone on which to build a set of standards and guidelines for statistical surveys. Basing their work in large part on FCSM Working Paper 31, described above, the Committee proposed the set of standards that were approved by FCSM and then adopted by OMB and issued in 2006 as Statistical Policy Directive 2 (OMB 2006).

A.4 Increasing Use of Nonstatistical and Integrated Data for Statistical Purposes (2011-2016)

Several OMB policy memoranda in the last decade provided support and led to the use of nontraditional sources and integration of nontraditional and traditional data sources to produce new federal statistical information products. Some examples include:

- OMB M-11-02, “Sharing Data While Protecting Privacy,” (OMB 2011) directs “agencies to find solutions that allow data sharing to move forward in a manner that complies with applicable privacy laws, regulations, and polices.” This includes “seeking ways to facilitate responsible data sharing for the purpose of conducting rigorous studies that promote informed public policy decisions.”
- OMB M-13-13, “Open Data Policy—Managing Information as an Asset,” (OMB 2013) was issued “to help institutionalize the principles of effective information management at each stage of the information’s life cycle to promote interoperability and openness.”
- OMB M-14-06, “Guidance for Providing and Using Administrative Data for Statistical Purposes,” (OMB 2014) “encourages Federal departments and agencies to promote the use of administrative data for statistical purposes and provides guidance in addressing legal and policy requirements for such uses, including the need to continue to fully protect the privacy and Confidentiality afforded to the individuals, businesses, and institutions providing the data.”
- OMB M-15-15, “Improving Statistical Activities through Interagency Collaboration,” (OMB 2015) “strongly encourages the Federal statistical agencies and units, and their parent Departments, to build interagency collaboration that will help the Federal statistical community more effectively meet the information needs of the 21st century.”
- OMB Circular A-130, (2016) “Managing Information as a Strategic Resource,” which establishes general policy for information governance, acquisitions, records management, open data, workforce, security, and privacy, was revised in 2016 to ensure consistency with these policy memoranda. In particular, these revisions clarified agencies’ information security responsibilities to align with the open data policy outlined in OMB M-13-13.

A.5 Plotting a Course for Federal Statistics (2015-2017)

In 2015, the Committee on National Statistics (CNSTAT) established a panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation “to conduct a study to foster a paradigm shift in federal statistical programs that would use combinations of diverse data sources from government and private-sector sources in place of a single census, survey, or administrative records source.” This CNSTAT panel released two reports in 2017.

In its first report, *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*, the CNSTAT panel acknowledged the increasing efforts over the previous decade to incorporate an ever-widening range of data sources into federal statistics, while noting that much work is still needed to achieve an effective shift from the single-survey paradigm to a multiple-source paradigm (CNSTAT 2017). Much of the report is dedicated to describing the range of potential sources that are and could be exploited to increase the quality and/or cost-efficiency of federal statistics, and/or to reduce response burden on the public. The first two recommendations are for federal statistical agencies to undertake systematic reviews of their existing statistical portfolios to evaluate the potential benefits and risks of using administrative and private sector data sources. To support these reviews, and to enable transparency with data users, the report advocates for the development of a data quality framework that contextualizes the strengths and weaknesses of different sources and approaches to their incorporation. It envisions a systemwide effort, in collaboration with academia and industry, to develop this framework and use it to create standards for evaluating integrated data product's fitness for various uses.

The CNSTAT panel's second report, *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*, expands on the first report's call for the systematic assessment of data quality, examining existing quality measurement approaches and identifying some desired features of an updated quality framework (CNSTAT 2018). It reviews the well-developed Total Survey Error framework that is usually applied to one data collection at a time, noting that agencies have built up careful protocols around this framework for understanding and reporting data quality. The panel recommends that federal statistical agencies adopt a broader data quality framework for statistical information that captures multiple dimensions of data quality, not just the accuracy dimension on which the Total Survey Error framework focuses. In particular, it identifies timeliness and granularity as two important dimensions to include. At the same time, the panel notes the need for each dimension to be deep enough to include pertinent aspects of nonstatistical and integrated data as well as survey data. The panel also notes the need to include well-developed treatments of data linkage errors and their potential effects on the resulting statistics. Recommendation 6-2 (p. 127) states that a useful framework for using alternative data sources "... should outline and evaluate the strengths and weaknesses of alternative data sources on the basis of a comprehensive quality framework and, if possible, quantify the quality attributes and make them transparent to users." The panel states further that "Agencies should focus more attention on the tradeoffs between different quality aspects, such as trading precision for Timeliness and Granularity, rather than focusing primarily on Accuracy" (CNSTAT 2018).

During 2016 and 2017, a separate effort to envision a new paradigm in federal statistics was also underway. The Commission on Evidence-Based Policymaking was established by the U.S. Congress (Evidence-Based Policymaking Commission Act of 2016 (Pub. L. No. 114-140)) and charged with developing a strategy to increase the availability and use of data to build evidence about government programs, while protecting privacy and confidentiality. Like the CNSTAT panel, the Commission focused substantial attention on the potential use of administrative data. Some of its 22 recommendations can be applied towards improving the federal statistical agencies' access to such data, while safeguarding the privacy of individuals and organizations that provide it. But the Commission also encouraged improved curation of data as it is being collected, with the maintenance of metadata to enable potential secondary users to assess and report on its fitness-for-use. The Commission's final report "envisions a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy." Throughout the report, the Commission emphasizes the importance of producing high quality data to enable evidence that supports policymaking (Commission on Evidence-Based Policymaking 2017).

A.6 Federal Efforts to Support Expanded Use of Nonstatistical Data (2017-present)

In response recognition to the Commission's recommendations, in the Fall of 2017 Congress began to consider legislation to increase the effective use of federal data to inform policymaking. On December 21, 2018 the Foundations for Evidence-Based Policymaking Act of 2018 was enrolled and on January 14, 2019, the Evidence Act was signed into law (Pub. L. No. 115-435, 132 Stat. 5529, 2018). The Evidence Act addressed 11 of the Commission's 22 recommendations and called for an Advisory Committee on Data for Evidence Building to continue the conversation on others. It articulated a presumption that data assets held by federal agencies will be made accessible to statistical agencies and units, while requiring those statistical agencies to expand the access they provide to data for evidence building. To advance data quality, it required federal agencies to designate Statistical Officials to advise the Department on statistical policy, techniques, and procedures. In memorandum M-19-23, OMB elaborated the responsibilities of the Statistical Officials to include serving as the "agency champion for data quality to ensure data Relevance (*e.g.*, by validating that data are appropriate, accurate, objective, accessible, useful, understandable, and timely), harness existing data (*e.g.*, by identifying data needs and reusing data if possible), anticipate future uses (*e.g.*, by building interoperability of data from its inception), and demonstrate responsiveness (*e.g.*, by improving data collection, analysis, and dissemination with ongoing input from users and stakeholders)" (OMB 2019c, 30). The Evidence Act also contains several provisions aimed at improving the quality of federal data assets, by requiring agencies to "develop and maintain a strategic information resources management plan that . . . implements a process to evaluate and improve the timeliness, completeness, consistency, accuracy, usefulness, and availability of open Government data assets" (Pub. L. No. 115-435, Title II, Sec. 202(c), 2018, 4174-8).

Meanwhile, the executive branch has also taken its own actions to promote the expanded use of existing federal data assets for evidence building. The President's Management Agenda "Modernizing Government for the 21st Century," launched in March of 2018, identified "Leveraging Data as a Strategic Asset" as a Cross-Agency Priority Goal. To pursue this goal, OMB developed the Federal Data Strategy; among the 10 governing principles it announced in June 2019 is "Ensure Relevance: Protect the quality and Integrity of the data. Validate that data are appropriate, accurate, objective, accessible, useful, understandable, and timely" (OMB 2019b). This principle is aligned with the Statistical Official's responsibility identified above. One of the 20 Action Items in the 2020 Action Plan is to "Develop Data Quality Measuring and Reporting Guidance" (which this report will, in part, fulfill).

New OMB guidance on the Information Quality Act also speaks to the emerging use of multiple data sources. In April 2019, OMB released memorandum M-19-15, "Guidance on Improving Implementation of the Information Quality Act." This updated guidance acknowledges that the CNSTAT recommended the use of a comprehensive quality framework that includes evaluating and documenting the timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy and confidentiality, transparency and interpretability, and granularity of each data source that is used to increase the integrity of analyses based on administrative data and data combined across two or more sources. OMB M-19-15 notes that The Foundations for Evidence-Based Policymaking Act of 2018 codifies these concepts as agency responsibilities. Because of the growing secondary use of data for purposes other than the reason the data were originally collected, OMB M-19-15 calls for the development of procedures for documenting and disseminating information on the quality of administrative data that have the potential to be used for statistical purposes, emphasizing the fact that the documentation should be sufficient to allow data users to determine the fitness-for-purpose of the data for use in secondary analysis (OMB 2019a).

A.7 ICSP and FCSM Products (2017–present)

In 2017, the FCSM established a Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources to support the identification of best practices around data quality measurement and reporting for integrated data products. The Working Group and the Washington Statistical Society co-sponsored several public workshops to gather information. The first three workshops focused on the quality of input data, data processing, and statistical outputs; additional workshops focused on the quality of geospatial data, transparent reporting of meta-data, and sensitivity analyses followed. They provided examples of the variety of ways agencies have integrated new data sources into their statistical programs, creating a variety of potential benefits such as cost savings and efficiency in production, increased timeliness and granularity, and reductions in measurement error. More information about the first three workshops can be found in JPSM report, *Findings from the Integrated Data Workshops Hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society* (Brown *et al.* 2018) sponsored by the Economic Research Service of the U.S. Department of Agriculture.

These potential benefits are weighed against the many challenges that are faced when using nonstatistical sources and integrating multiple sources. From an input perspective, assessing the quality of alternative data sources is similar in many dimensions to assessing traditional survey data concerns over coverage and representativeness, coherence and consistency over time, reporting errors, and missingness can be mapped to traditional survey error concepts relatively straightforwardly. There are fewer parallels for less structured and unstructured data sources, where development of ontologies and models is necessary to use the data in statistical estimation. With many kinds of external data sources, the statistical agency has less control over and less information about the data than in the traditional survey context; transparency around the original purpose of collecting the data was emphasized.

The various processes employed in the integration of data sources (*e.g.*, record linkage, data fusion, and systems for harmonizing variables) introduce additional data quality challenges. The workshops highlighted the importance of providing transparency to data users about the techniques used to integrate sources, the assumptions that underlie them, and the biases these might create in resulting integrated estimates and data products. Approaches for measuring the quality of integrated sources tend to vary based on the available tools. For example, a high quality “truth deck” can be extremely useful for determining the quality of a linked data product. When no verification data are available, sensitivity analysis can provide valuable information about the importance of modeling assumptions on the statistical output. New methods for sensitivity analysis continue to be developed and may be able to be adapted by statistical agencies.

To complement the information from the workshops, the SOI Division of the Internal Revenue Service sponsored a report by Mathematica Policy Research that examined data quality frameworks and standards used outside the United States by national statistical offices and international organizations, including the European Statistical System and a selection of individual European countries, Canada, Australia, the OECD, and the IMF. As in the workshops, 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. In addition, data quality is expressed in each of these quality frameworks as multi-dimensional, with recognized trade-offs among the dimensions and with the accuracy dimension often given the most attention.

The ICSP has established a set of principles on which to base the development of a data quality framework and standards for agencies in the United States. They note that:

All data have potential errors, and errors can be compounded when data from different sources are integrated to produce statistical estimates. Poorly estimated or overextended statistics can misguide decision makers, leading to costly consequences. Federal statistical agencies must ensure that nonstatistical or integrated data sources result in quality statistics and clearly, meaningfully, and effectively communicate the limitations of those statistics, so they are used wisely (ICSP 2018).

The ICSP principles place a high priority on using the best quality source data, while recognizing the value of granularity and timeliness as aspects of that quality. They emphasize the importance of reporting transparently to data users on the strengths and limitations of disseminated data, including the assumptions and uncertainties that may underlie them, in a way that is relevant to the expected applications of the data. Finally, they state that:

Agencies should work to adopt common language and framework for reporting on the quality of data sets and derivative information they disseminate. The focus should be on providing information that will meet user needs, which may vary by product and agency. Drawing on industry standards, where they exist, will improve interoperability of federal and nonfederal data (ICSP 2018).

References

- Atkinson D, Schwanz D, Sieber WK. 1999. Reporting Sources of Error in Analytic Publications. Seminar on Interagency Coordination and Cooperation. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 28). 329–341.
- Brown A, Abraham KG, Caporaso A, Kreuter F. 2018. Findings from the Integrated Data Workshops hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society, available at https://nces.ed.gov/fcsm/pdf/Workshop_Summary.pdf.
- CNSTAT, National Academies of Sciences, Engineering, and Medicine. 2017. Innovations in federal statistics: Combining data sources while protecting privacy. National Academies Press.
- CNSTAT, National Academies of Sciences, Engineering, and Medicine. 2018. Federal statistics, multiple data sources, and privacy protection: Next steps. National Academies Press.
- Commission on Evidence-Based Policymaking. 2017. The Promise of Evidence-Based Policymaking, available at <https://www.cep.gov/cep-final-report.html>.
- FCSM. 1987. An Error Profile: Employment As Measured By The Current Population Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 3).
- FCSM. 1988. Quality in Establishment Surveys. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 15).
- FCSM. 1990a. Survey Coverage. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 17).
- FCSM. 1990b. Data Editing in Federal Statistical Agencies. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 18).
- FCSM. 2001. Measuring and reporting sources of error in surveys. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 31).
- ICSP. 2018. Principles for Modernizing Production of Federal Statistics, 2018. Available at <https://nces.ed.gov/fcsm/pdf/Principles.pdf>.
- OMB. 2002a. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies, 67 FR 8451.
- OMB. 2002b. Federal Statistical Organizations' Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Disseminated Information, 67 FR 38466.
- OMB. 2006. Standards and Guidelines for Statistical Surveys, 2006. 71 FR 55522. Addendum available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/directive2/final_addendum_to_stat_policy_dir_2.pdf.
- OMB. 2011. Memorandum M-11-02. Sharing Data While Protecting Privacy, available at <https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2011/m11-02.pdf>.
- OMB. 2013. Memorandum M-13-13. Open Data Policy-Managing Information as an Asset, available at <https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf>.

OMB. 2014. Memorandum M-14-06. Guidance for Providing and Using Administrative Data for Statistical Purposes, available at <https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2014/m-14-06.pdf>.

OMB. 2015. Memorandum M-15-15. Improving Statistical Activities through Interagency Collaboration, available at <https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2015/m-15-15.pdf>.

OMB. 2016. Circular A-130. Managing Information as a Strategic Resource, available at <https://www.cio.gov/policies-and-priorities/circular-a-130/>.

OMB. 2019a. Memorandum M-19-15. Guidance on Improving Implementation of the Information Quality Act, available at <https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf>.

OMB. 2019b. Memorandum M-19-18, Federal Data Strategy - A Framework for Consistency, available at <https://www.whitehouse.gov/wp-content/uploads/2019/06/M-19-18.pdf>.

OMB. 2019c. Memorandum M-19-23, Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance, available at <https://www.whitehouse.gov/wp-content/uploads/2019/07/M-19-23.pdf>.

Pub. L. No. 114-140. 2016. Evidence-Based Policymaking Commission Act of 2016

Pub. L. No. 106-554, § 515(a). 2000. Information Quality Act.

Pub. L. No. 115-435, 132 Stat. 5529. 2018. Foundations for Evidence-Based Policymaking Act of 2018.

Appendix B. Accuracy and Reliability of Integrated Data

B.1 Overview

The accuracy and reliability of data are frequently the focus of data quality evaluations. This topic deserves a more robust discussion for interested readers, especially considering the importance of accuracy and reliability when data are released as official statistics or used as influential inputs to policies and decisions (OMB 2019). Many factors that affect accuracy and reliability and their effects have been well-described as error sources for surveys in the Total Survey Error (TSE) paradigm (see, *e.g.*, Groves 1989, FCSM 2001). However, there are additional threats unique to the accuracy and reliability of integrated data, including linkage error and modeling error. More generally, threats to accuracy and reliability for integrated data include threats of the input data sources, the processing steps, and the resulting data outputs. Approaches beyond those used for survey data will be needed to examine and convey uncertainties as new types of integrated data products and outputs are developed.

Given the variety of statistical and nonstatistical data sources, integration methods and data products, the best approaches for measuring and reporting error for integrated data will depend on the data product under evaluation. Potential approaches include: a) direct comparisons to an external gold standard; b) benchmarking with known subject-specific and other relevant information; and c) conducting sensitivity analyses, such as evaluations of critical processing and analysis decisions, generally applied to modeling assumptions used in an analysis (Goldsmith 2005). Established high-quality national statistical data for use as gold standard data and benchmarking will continue to be critical for assessing the quality of new data products.

B.2 Precise and Unbiased Information

In sample surveys, accuracy can be formally described as the closeness of a sample-based statistic, such as a sample mean (\bar{x}) (or sample median, proportion) to its population-level counterpart, such as a population mean (μ). Groves *et al.* (2009) divides the errors encountered in sample surveys into two categories: errors of measurement, which affect the accuracy of any single observation and errors of representation, which refer to factors that diminish how well a sample of units “represent”—portray, capture, or measure—the characteristics, conditions, experiences or behaviors of the population. Whereas these concepts are traditionally presented in the context of sample surveys, they apply broadly to all data that are used to measure or reflect a characteristic of some population of interest. Indeed, Zhang (2012) extends the concepts to administrative and integrated data.

The effects of such errors may be usefully divided into two, broad subcomponents: a systematic component in which the measured data deviates from the true values in a nonrandom way and a nonsystematic component which introduces random variation in any set of estimates. These components of error map into two components of accuracy: statistical unbiasedness and statistical precision.

Statistical unbiasedness refers to a condition where “the expected value of an estimate of a characteristic is equal to the true population value” (CNSTAT 2018, 40). The phrase “expected value of an estimate” refers to a value that would emerge if the population were sampled and re-sampled many times, the estimate calculated for each sample, and an average (expected value) of all sample estimates calculated from all the possible samples. A broader concept of unbiasedness is also found in the dimension scientific integrity, within the data quality framework’s integrity domain (OMB 2008, OMB 2014, ICSP 2018).

Statistical precision, defined as the inverse of the variance, indicates the variability of an estimate. Statistical precision is related to the errors, or threats, that affect variance and standard errors, including sampling error and modeling error. The ICSP defines statistical precision as “a measure of how close two or more measurements of the same statistic are to each other” (ICSP 2018). Estimates with high statistical precision have lower variance and are more accurate than those with low statistical precision. Like unbiasedness, precision has constructs in other disciplines that differ from its statistical use here. For example, numerical precision is the number of digits in a number; precision can also indicate, resolution (or granularity), the smallest interval measurable by the scientific instrument (*i.e.*, our data collections).



B.3 Threats to Precision and Sources of Bias

This section describes the various kinds of error that can harm the precision of estimates and create bias. It includes threats that affect statistical surveys like measurement, sampling, nonresponse, coverage, and processing errors, as well as threats unique to integrated data like linkage, harmonization, and modeling errors. We also discuss some additional threats specific to geographic data.

Measurement error: In the framework of TSE, the accuracy of a survey statistic can be affected by two sets of errors, which are classified as errors of representation and errors of measurement in the context of a survey life cycle (Groves *et al.* 2009, 48). There are two definitions of measurement error. According to one definition, measurement error is defined as the “difference between the observed value of a variable and the true, but unobserved, value of that variable” (FCSM 2001, 1-6). Alternatively, the term “measurement error” may refer to the variance (or standard deviation) of that difference. In a survey, measurement error typically occurs when a respondent provides inaccurate information, for example, if the respondent misunderstands the survey question or if they have poor recall of the items they are being asked about.

Frequently measurement error is conceived as adding to the variability (imprecision) of an estimate rather than contributing to a biased estimate, which is a scenario in which the variance (or standard deviation) of the error can serve well as a measure. However, the literature on TSE conceived of measurement error broadly and allows for such error to affect either the variance or the bias of an estimate (Groves, 1989, 17). In the context of survey data, measurement error comes from four primary sources in survey data collection: the questionnaire, the data collection method, the interviewer and the respondent.” (FCSM 2001, 1-6). Examples of mode effects include differences among in-person face-to-face interviews, telephone interviews, and web surveys; within web surveys, there is also the possibility of mode differences in how respondents respond to an interview administered on a desktop computer, a laptop computer, a tablet, or on a handheld phone. The questionnaire or survey script can contribute to measurement error if the information requested is not clear to respondents, or if the skip patterns are unclear. Finally, the respondent can add to measurement error by reporting incorrect responses—either because the respondent does not want to reveal a sensitive piece of information or attempts to give socially desirable answers rather than the correct answer.

For a given data collection, agencies may report on various aspects of measurement error in different publications. Although ideal, there may not be a single, comprehensive publication by an agency that collates all the work the agency and others have done to study the data collection’s measurement error, particularly if data collections and assessments are ongoing. However, it could be helpful for an agency publication that contains its estimates to include: (a) a brief description of measurement error and its possible effects, and (b) citations to studies that the agency (or others) have conducted on measurement error that are pertinent for the data collection at hand.

Agencies employ a variety of approaches to identify and address possible sources of measurement error before an instrument goes into the field. For example, an agency may test the data collection instrument using cognitive interviewing to identify possible interpretations of questions and evaluate question-response patterns (*e.g.*, NCHS 2020), which allows the instrument to be revised and perhaps re-tested before implementation.

Similar factors can contribute to measurement error in nonsurvey data. For secondary uses of nonstatistical data, the measurement error can differ among its original purpose and its intended uses. Further, the importance of measurement error and the efforts to reduce it may differ between its original purpose and its intended uses. Importantly, for data originally collected for administrative or other purposes, the primary documentation necessary for detecting and remedying these errors may not be available.

Sampling Error: Sampling error has long been associated with survey data and the probability-based surveys that generate them. Obtaining data for an entire population is expensive, if even possible. However, a consequence of using survey-based data is that samples have sampling error attached to them. Threats to accuracy for integrated data that combine surveys or combine surveys with other types of data will include the sampling error of the respective surveys.

Perhaps the most common definition for sampling error is that such error is the difference between a sample-based statistic and the (unknown) population value. For example, when using the sample mean \bar{X} as an estimate for the population mean μ , we can represent the sampling error as the difference $\varepsilon = \bar{X} - \mu$. By this definition, if the sample is not a subset but the full set of the population—if the sample is tantamount to a census—then the population value would be known (apart from other, nonsampling errors, described below) and there is no sampling error. This symbolic definition $\varepsilon = \bar{X} - \mu$ is captured in the textual definition in Statistical Directive 2 which states that “Sampling error is the error associated with nonobservation, that is, the error that occurs because all members of the frame population are not measured” (OMB 2006, 34).

A second definition of sampling error builds upon the first definition. Using the first definition, sampling error *per se* cannot be measured; that is, $\varepsilon = \bar{X} - \mu$ cannot be measured because that calculation would require knowledge of the (unknown) population parameter μ . However, statistical procedures are available to measure the variability of sampling error in two ways—the variance of sampling error $V(\varepsilon)$ and the standard error of sampling error $SE(\varepsilon)$, where the latter is the square root of the former. The term sampling error sometimes refers to the variability of sampling error, usually measured by $SE(\varepsilon)$. This second meaning for sampling error is captured in an FCSM working paper, which states: “Sampling error refers to the variability that occurs by chance because a sample rather than an entire population was surveyed” (FCSM 2001, 1-5, emphasis added).

Thus, by one definition sampling error is ε and by another it is $SE(\varepsilon)$. Context usually indicates which meaning is intended. In fact, discussion of the many types of errors—not just sampling error—often (implicitly) involves the variability of the error rather than the error itself. The term standard error may refer to the standard error for a sample-based estimate, such as the standard error of a sample mean, $SE(\bar{X})$ and the phrase standard error of the sampling error, as measured by $SE(\varepsilon)$, may not even appear. However, the two standard errors are equivalent. Thus, although sampling error and its variability may seem removed for common statistical practice, these concepts tie directly with how much variability there is in the estimates that users of the statistics care about directly: the sample-based statistics for means, totals, proportions and other statistical quantities.

Due to the random sampling methodologies that well-designed probability-based surveys employ, sampling error is unbiased. Even though individual samples may differ from the overall population, because the process is unbiased, differences are not consistent or directional. Instead sampling error threatens statistical analysis by reducing precision. Specifically, a large sampling error and large $SE(\epsilon)$ reduces the ability to use sample data to learn about the characteristics, conditions, experiences or behaviors of a population. That is, a large sampling error hinders statistical inference, where “inference” refers to the use of a sample-based statistic such as \bar{X} as an estimate of the population parameter μ , within a specified degree of precision.

Agencies can address sampling error at two different stages of the production process—the design stage and the fielding stage. At the design stage, agencies can choose, with considerable foreknowledge, how much variability or sampling error $SE(\epsilon)$ will be allowed in a survey. For example, they may reduce $SE(\epsilon)$ by collecting a large sample size, so long as budgetary resources are available. The choice of the sample design will also impact sampling error, for example choosing a stratified sample, rather than a simple random sample. High professional standards and careful implementation are needed so that the desirable statistical properties that a probability-based sample offers in principle are achieved in practice. If executed well, a probability-based survey can be used to produce estimates of sampling error that are meaningful, that is, the estimates measure what they are designed to measure. Note that there is no inherent methodology trade-off between reducing sampling error and achieving other statistical goals, that is, a lower $SE(\epsilon)$ does not worsen, say, other dimensions of quality, except as they relate to cost. (FCSM 2001).

Sampling error is typically captured through the standard error. As noted by the FCSM (2001, 1-6), “(f)or any survey based on a probability sample, data from the survey can be used to estimate the standard errors of survey estimates.” Its measurement is often carried out via closed-form formulae that account for the sample design of the data. However, it may also be calculated through numerical methods such as balanced repeated replication (see, *e.g.*, Kish and Frankel, 1970). CNSTAT (2017) describes sampling errors as measures of precision for survey estimates (p. 112) that provide the ability to measure the nature and extent of uncertainty in inferences to the full population (p. 89). Sampling error for a probability-based survey is measured and reported so that users can use the statistical information to support inference appropriately. Alternatively, if sampling errors are not reported, statistical tools, such as replicate weights in microdata files, by which researchers can generate their own estimates of sampling error, can be provided. By either method, users are enabled to conduct research, use the estimates, and make inferences appropriately. Sampling errors can be reported using coefficients of variation, standard errors, or confidence intervals. The suitability of each concept depends on the type of variable.

Based on the definitions above, it may seem at first that sampling error has no bearing on census data or, for that matter, on any data covering a population universe, or 100 percent of the population. However, another perspective is that a census can be interpreted as a type of a sample for some uses. Underlying this perspective is the argument that: “A census shows what resulted from this combination [of chance and underlying social and economic cause systems] at a certain time in the past, but any generalizations that are not restricted to a particular date and place must recognize the fact that some other population might have resulted, and must in fact be expected to arise in the future from the same underlying causes.” (Deming and Stephan, 1941, 45, emphasis added). The issue of how to interpret census data, or other universe data such as vital records on births and deaths as a concrete example, arises when using such data to make statistical comparisons, forecasts, and hypothesis tests involving regions, groups or time periods and answer questions about whether such comparisons differ by “more than some level of natural fluctuations” (Brillinger, 1986, 693).

Nonresponse Error and Missing Data: A definition of nonresponse error for survey data is “an error of nonobservation reflecting an unsuccessful attempt to obtain the desired information from an eligible or sampled unit,” where the unit is eligible for inclusion in the survey (FCSM 2001, 1–6). In this context, a unit can refer to a sampled individual, household, establishment or other entity. Unit nonresponse occurs when no information is obtained from a sampled unit. When a sampled unit provides information or answers to some, but not all questions, the missing responses are identified as item nonresponse. In addition to possible increases in bias, described below, nonresponse can decrease statistical precision through decreased sample sizes or, if the missing data are imputed, through added variability due to imputation.

In nonsurvey data, unit missingness can be considered to be an analog to unit nonresponse in a survey; some members of the set of units chosen, targeted, or scheduled for inclusion in the dataset are not, for whatever reason, included. Although unit missingness would be called nonresponse in a survey setting, items may be missing from nonsurvey sources for various reasons. Sensor and satellite data may be missing observations due to failing equipment, disruptions in signal detection, errors in recording or transmitting received signals, and such. Administrative sources that rely on voluntary reporting can miss the observations of population members who have not provided data. Each of these sources of missingness may operate on the item level for nonsurvey data, for example if a voluntary reporter provides information on a subset of requested elements. As secondary uses of data are often not considered by the data collector, efforts to track and reduce these sources of missing data may not be as strongly pursued as for statistical data, collected explicitly for statistical use, and the efforts that are made to track and reduce missing data may not be systematically documented.

Depending on missingness patterns, integrating data can mitigate or exacerbate the impact of missingness. Integrated data may have some observations dropped due to format incompatibilities or other logistical issues.

If item or unit nonrespondents in a survey differ systematically from respondents (*i.e.*, the nonrespondents are not missing at random), then the survey-based estimated value systematically deviates from the population parameter that the statistician seeks to measure, resulting in nonresponse bias. Missingness in nonsurvey and nonstatistical data has an analogous effect: if it is nonrandom with respect to the variable being measured, it results in bias. If nonresponse or missingness is completely random, then (this type of) missingness does not cause bias. However, the conditions under which missingness is completely at random may be rare; instead, certain types of units may be more likely to exhibit missing data than other types of units. Potentially, data on observable factors may (fully) account for this type of missingness. If so, then statistical techniques exist that can be used in estimation to eliminate any bias. Even if bias cannot be eliminated entirely, it may be satisfactory to use these techniques, together with at least some observable factors, to reduce potential bias.

In “A Systematic Review of Nonresponse Bias Studies in Federally Sponsored Surveys” (FCSM 2020), the FCSM examined methods used in 165 eligible nonresponse bias studies conducted for federally sponsored surveys after 2006. For this report, nonresponse bias methods were grouped into four categories: (1) benchmarking, (2) comparisons to external data, (3) studying variation within the respondent set, and (4) comparing alternative post-survey adjustments. The authors report that comparisons of survey estimates to external data were the most commonly used method to assess nonresponse bias for establishment surveys whereas studying variations within the respondent set was the most commonly used method for household surveys. Such studies may be used to develop unbiased or less-biased estimates; alternatively, results can be reported to provide data users with important information about nonresponse for their inferences from the data.

FCSM Working Paper 31 (2001, 4-18 and 4-19) describes several reporting practices that agencies can use to inform users about the potential effects of nonresponse and missing data. In analytic reports, they may publish the response rates or missingness rates at the unit and item levels for various populations and subpopulations, identify the methods employed to adjust nonresponse or missingness, and report on studies that have been conducted to measure bias. In technical reports, they may provide additional details about the procedures used to mitigate nonresponse and missingness, report on evaluations of these mitigation procedures, and describe steps taken to identify and address the underlying causes of nonresponse or missingness.

For integrated data, sources having high rates of nonresponse or missingness may be excluded from the integrated data, especially if such missingness exacerbates other threats to data quality. Decisions about sources with missing data must weigh the inherent trade-offs between potential losses in accuracy and reliability against potential gains in other quality dimensions such as relevance.

Coverage Error To survey statisticians, the population of individuals, households, businesses or entities that the statistician aims to represent in the sample is the target population. In contrast, the population from which selected members are drawn for the survey sample is the frame population. Ideally, the target population and the frame population are the same, but differences can emerge. Statistical results are necessarily based on data at hand from a frame population from which a representative sample is drawn (survey data) or administrative data are collected (nonsurvey data). In either case, a disparity between the frame and target population means that data products do not fully represent the population of interest.

FCSM Working Paper 31 identified that “The source of coverage error is the sampling frame itself” (FCSM 2001, 1-6). Specifically, coverage error occurs when the sampling frame differs from the target population (*i.e.*, there is not a one-to-one match between the frame and the target population) (Groves *et al.* 2009, 55). Substantial coverage errors affect the ability of the user to make inferences about the target population.

Undercoverage occurs when the frame used for sampling does not include all the units or people in the target population. Conversely, overcoverage occurs when the sampling frame includes more units or people than the target population (FCSM 2001, 1-6) In an address-based frame some housing units may be missed during enumeration and will result in undercoverage, (*e.g.*, multiple housing units may not be apparent in subdivided houses). In contrast, if the frame is based on telephone numbers, households with two telephone numbers will be duplicated on the frame resulting in overcoverage.

Nonsurvey data contain errors that can be considered to be coverage errors. In nonsurvey data, the universe from which the data are actually drawn (a set that is the counterpart to the frame population) may not match the universe that data-based estimates are intended to represent (the target population). In nonsurvey data, it may not be as common as in survey data to define and distinguish between a study’s target and actual (*i.e.*, frame) universes formally, but such definitions should be made explicit in documentation of nonsurvey data.

In many secondary uses of nonsurvey data for statistical purposes, coverage errors may be especially common. Whereas frame populations in the survey context are usually chosen carefully to match the target population during the survey’s design stage, secondary uses of data usually exploit data sets whose frames were selected and developed independently from the secondary use. In addition, when data sets having different frames are integrated, the impact of coverage errors may be exacerbated. On the other hand, careful treatments may diminish coverage errors by optimally using different data sets to

measure different segments of the target population. For example, Schenker, *et al.* (2002) combine data from the National Health Interview Survey and the National Nursing Home Survey to estimate the overall prevalence of health conditions among the elderly (living in either households or nursing homes). It is particularly useful to have at least one data set that covers the entire population for a set of key covariates when combining sources this way. For example, the National Postsecondary Student Aid Study links to Federal Student Aid records for information on family ability to pay and the receipt of federal student aid; however, the Federal Student Aid records lack this key information for students whose families did not complete the Free Application for Federal Student Aid (also known as the FAFSA) (Wine *et al.* 2018).

For statistical data, designed and collected for statistical purposes, such as survey data, it may be possible to consider specifically how to limit coverage error by carefully specifying the target population and then collecting data from it. For secondary use of data originally collected for another purpose (nonstatistical data), it is not possible to change the data or specify its design. However, the threat of coverage error can be addressed in an indirect way by being aware of possible discrepancies between the frame population(s) used in the set(s) of data and the target population of interest. Such awareness reduces the chances of adopting statistical conclusions without caveats. Reporting on possible disparities between a target and a frame population for both statistical and nonstatistical data is one way to address coverage error; in so doing, data products based on the frame population are not improperly attributed to the target population.

There can be uncertainty about the magnitude of coverage error, that is, the extent to which the frame population from which the nonsurvey records were drawn does not fully capture the desired target population. For example, some administrative datasets track the participants in a particular program; we may consider these participants to be a subset of a larger frame population of people who were eligible to participate in the program. That is, not all people who are eligible to participate (the frame population) become actual participants. When there is unmeasured heterogeneity in the extent to which those eligible for a program were engaged to participate, it can result in ambiguity about how the frame population is best defined. Such ambiguity adds error to the mapping of the data to a target population. This issue highlights the importance of having good metadata and documentation for each data source. For example, the National Postsecondary Student Aid Study sample frame is matched to records at the Veteran's Benefit Administration to identify veterans in the sample frame to allow for oversampling. However, the veterans are limited to those who apply for benefits from the Veteran's Benefit Administration (Wine *et al.* 2018).

Processing Error: "Processing error occurs after the survey data are collected, during the processes that convert reported data to published estimates and consistent machine-readable information. Each processing step . . . can generate errors in the data or in the published statistics." (FCSM 2001, 1-6) As is the case with other types of errors, these errors occur when a survey value is different from the true response. Commonly cited types of processing error include data entry, coding, editing, imputation, and analysis (FCSM 2001, Czajka and Stange 2018, Eurostat 2020). The 2020 European Statistical System Handbook for Quality Reports distinguishes between errors that occur in processing of the microdata into machine readable formats (see processing error examples above) and those errors that involve mistakes in implementing procedures (Eurostat 2020).

FCSM Working Paper 31 (FCSM 2001, 7-1) states "Data entry errors occur in the process of transferring collected data to an electronic medium." This FCSM report acknowledges that in data entry, errors can be "considerably reduced" as the technology used for data entry increases from paper and pencil to automated scanning, to computer assisted data collection, and

to web collection techniques; it also cautions that “data entry errors occur even with technologically advanced techniques” (FCSM 2001 7-1).

On the topic of coding edits, FCSM Working Paper 31 (FCSM 2001) indicates that “Most surveys require some type of pre-edit coding of the survey returns before they can be further processed in edit, imputation, and summary systems.” (FCSM 2001, 7-4). At this stage in processing—unit response— coding is used to indicate whether enough information was provided for a usable case. Item response coding is used to code short answer and open-ended responses into categorical responses. By the very nature of the fact that the coding is conducted by people and judgement is involved, the possibility of errors in coding is real. Automated coding of open-ended responses also serves to decrease coding errors (FCSM 2001).

Editing, and thus editing errors, occur at different points throughout processing (*e.g.*, “from manual editing prior to a machine edit or from the manual correction of computer-generated error flags” (FCSM 2001, 7-6). Although editing is intended to improve the quality of the data, Working Paper 31 reported research showing that over-editing can add more error than it eliminates, so finding the right balance is important.

FCSM (2001) also discussed imputation error, noting that when imputation is used to replace missing or incorrect variables, the missing or erroneous data are replaced with values generated by an assumed model for the nonrespondents’ data. For example, in hot deck imputation, the problematic or missing values are replaced by reported values from other units in the survey. However, if the models are incorrect, the imputation process will introduce error. FCSM (2001) highlighted the point that “even if the imputation models are reasonable, imputation often attenuates measures of association among variables.” As a result, the authors stressed the importance of documenting imputation procedures and including imputation flags for imputed cases to allow future analysts to remove the existing imputations and compute their own.

Although most of the attention on processing error has been given to survey data and other data within the control of the statistical agency, less will be known about processing errors for external nonstatistical data obtained for statistical purposes, either on their own or for integrated data products. It is likely that data processing will differ among external data providers and will change over time, so that processing errors for a particular data element could change with changes in data providers or their likely unknown processing steps, and these changes may remain unknown.

Linkage Error: Record linkage is the integration of two or more record level data sets using unique or common identifiers to produce a linked data file with information from each source for each record. Linkage errors decrease the accuracy of linked data, which can lead to increased statistical bias and standard errors in the resulting estimates. Linkage errors are typically described as Type I (false match) or Type II (missed true matches). The quality and type of identification information used to link cases has significant bearing on these error rates and the resulting accuracy of the linked data resource.

To facilitate the highest quality linkages, the ideal situation is to have a unique identifier, such as an individual’s Social Security Number (SSN), associated with all records in the data sources to be linked. But often, such an identifier is missing in one or more of the data sources, making accurate record linkage more challenging. For example, Massey and O’Hara (2014) used a combination of name, age, and birthplace as the core identifiers to match 1940 Census records with data from later Census years, given the lack of a unique identifier on the 1940 Census file.

The two general approaches to record linkage are deterministic and probabilistic matching techniques (Dusetzina *et al.* 2014). Deterministic matching algorithms generally compare identifiers or groups of identifiers across two or more sources of data and set criteria for matches on exact agreement between the identifiers. Deterministic matching algorithms tend to be rules based. Probabilistic matching algorithms generally calculate the probability or likelihood that two sets of records are correctly matched based on a sum of calculated probabilities for agreement and nonagreement for each identifier used in the linkage. The weighting procedures to calculate the probabilities typically follow the Fellegi-Sunter paradigm which is the foundational methodology used for record linkage (Fellegi and Sunter, 1969).

When linking with a unique identifier such as an SSN, the deterministic methods may work well due to the fact that exact matches are required. As a consequence, this approach tends to have fewer Type I (false) matches. But, if errors occur in the linking variable(s) (*e.g.*, due to recording errors) or if the linking variables are not unique, the number of Type II (true missed matches) can increase (Grannis *et al.* 2002). For example, Handwerker, *et al.* (2011) used Employer Identification Numbers (EINs) to link firm-level data on foreign-owned businesses from the Bureau of Economic Analysis with establishment-level data on employment from the Bureau of Labor Statistics. Since firms often use multiple EINs, and the firm-level data does not include all of each firm's EINs, the project augmented its deterministic matching approach with auxiliary identifiers, but still matched only 453 of the 500 manufacturers it studied.

Probabilistic record linkage may be better for linking data when linking identifiers may be subject to change over time and other reporting, measurement, or recording errors (Newcombe *et al.* 1959, Fellegi and Sunter 1969, Sayers *et al.* 2015, Harron *et al.* 2017a). As the term implies, the probability of records matching is provided, and a cut-off score is used for determining a match and nonmatch. Sometimes a second score is used to separate matches, possible matches, and nonmatches. The possible matches are then subjected to manual review. As the cut-off values for matches and possible matches increase, the likelihood of incorrectly linking records decreases (Type 1) while the probability of missing matches increases (Type II) (Krewski *et al.* 2003). The choice of optimal cut-off values is not straight forward at best and is often subjective. Manual review of record pairs and plotting the distribution of matched weights are useful tools (Blakely and Salmond 2002, Dusetzina *et al.* 2014). If a single cut-off is used to identify matches and nonmatches, then setting that cut-off value can be aided by conducting sensitivity analyses. Evaluation work (empirical and theoretical) should be conducted to compare recommendations for best practices associated with probabilistic record linkage based on typical uses of the data.

Harron *et al.* (2017a) provide three approaches to assessing the quality of record linkages: (1) applying the linkage algorithm to a subset of gold standard data to quantify linkage error, (2) comparing characteristics of linked and unlinked data to identify potential sources of bias, and (3) evaluating the sensitivity of results to changes in the linkage procedure, algorithm or thresholds. For example, the linkage of the National Hospital Care Survey with the NCHS National Death Index incorporated the three approaches: (1) a test deck/gold standard was created based on linking the SSN and other identifiers for verification, (2) the discharge status was compared with the linked data to assess whether there was bias for those discharged dead and not linked, and (3) match weights were assessed to determine if different thresholds would produce different results (NCHS 2019).

A recent paper by Resnick and Asher (2019) describes the most common methods used for estimating Type I and Type II errors in record linkage, focusing on a gold standard approach, a data dependent analysis, and a simulation. A recent linkage project at NCHS utilized aspects of these approaches to produce high-quality matches with a low degree of error (NCHS

2019). Match weight thresholds were determined by the lowest total error for Type I and Type II errors. Type I and Type II errors were estimated using the gold standard approach, specifically “Records that are matched by the probabilistic record linkage process but not the deterministic process are false positives [Type I]; those matched by the deterministic process but not the probabilistic process are false negatives [Type II]. From counts of these matches, Type I and Type II error rates can be determined” (NCHS 2019).

The error structure of the matching can be made available to the user or agency so that methods that account for the linkage errors can be incorporated into the data analysis when using the linked data file for research or estimation. For example, Lahiri and Larson (2005) incorporated matching error into regression analysis and Hof *et al.* (2017) incorporated matching error into a survival analysis.

The NCES National Postsecondary Student Aid Study is a complex, nationally representative cross-sectional study of students attending postsecondary institutions eligible for student financial aid from the federal government. The National Postsecondary Student Aid Study covers topics pertaining to student enrollment in postsecondary education, with a focus on how individuals and families finance postsecondary education. It includes a student survey as well as the collection of data from the institutions in which the study students are enrolled. Record linkage is used to add administrative information from the Department of Education’s Federal Student Aid Office, the National Student Clearing House, the College Board, the ACT organization, and the Veterans Benefits Administration. Technical documentation provides information about each data source, including the match rates for linkage to each source (Wine *et al.* 2018), and including match rates for different subpopulations of interest.

Although record linkage is performed to increase the relevance and accuracy/reliability of data through the combination of information from more than one data source, (see harmonization below), the linkage process affects the timeliness of the data and increases risks to coherence. Approaches for evaluation include:

- Estimate the linkage error using a gold standard (also known as a truth deck) approach, data dependent analysis, and/or a simulation approach
- Report the linkage error in documentation
- Provide probabilistic match weights so researchers can conduct sensitivity analyses with different threshold cutoffs

Harmonization Error: Harmonization, by definition, is the process of bringing something into consonance or accord (<https://www.merriam-webster.com/dictionary/harmonize>). For integrated data products, this means the process of mapping and synchronizing data derived from multiple sources into a coherent data file for analysis. Harmonization can be implemented using one source as a reference or gold standard or by using all sources to measure an underlying latent construct. Harmonization errors may increase statistical bias and decrease statistical precision of the resulting estimates.

Harmonization errors arise when elements in each data source have been defined, collected, or processed differently. Examples include the use of different classification systems, time periods or spatial supports, and the collection of different constructs, (*e.g.*, completed years of education versus earned educational certificates/degrees). Even when items are collected similarly, different levels of measurement error may require harmonization. Income data, for example, are often collected on population surveys and administrative records. However, this information can be collected for use at the individual, family

and/or household level and may be expressed relative to a federal threshold in administrative data collected for federal benefits. Income also changes over time and is subject to measurement error, depending on the collection method, intent of the collection, and specificity of income sources included in the collection instrument.

Efforts to align data elements from different sources within an integrated dataset or data collection can lead to harmonization errors. Harmonization error can arise from many situations, including lack of enough information about, and quality of, source data, and its impact may differ among integration methods and uses. Integrating disparate sources may entail projecting estimates across time, geography, or classification system. Holan (2018), for example, describes methods for spatial and spatial-temporal data available for different geographic units and time periods (also known as the “change of support problem”) that could be used to harmonize data with different support, including recent Bayesian models (Bradley *et al.* 2016).

Efforts to address measurement error in the source data may differ among sources and contribute to harmonization error. Survey questions measuring the same construct often differ among sources, particularly for items that are new or changing, such as sexual orientation and gender identity (see, *e.g.*, Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys 2016). Capturing similar information from administrative sources may be done in different ways. Harmonization of geographic information and other hierarchical structures can affect granularity. As a result, harmonizing responses among various data sources may affect the accuracy and reliability of the integrated data product, particularly if some categories are coarsened for the resulting product, such as combining categories of race and ethnicity or collapsing geographic codes. On the other hand, integration may improve estimation if sources with less measurement error can be identified and used for outputs from the integrated data product that otherwise would not have been available.

Jang (2018) describes data harmonization for the U.S. Scientists and Engineers Statistical Data System, which includes data from the National Survey of Recent College Graduates, the Survey of Doctorate Recipients, and the National Survey of College Graduates. Combining data from these input sources allows for nationally representative estimates of the whole science and engineering population, increases temporal coverage and provides information over time. However, changes in sample designs, collection protocols, variable naming conventions, formats, and other factors over time required consistent coding and editing procedures, including harmonized response rate calculations, imputations, weighting, and variance estimation.

Approaches for identifying, mitigating, and reporting harmonization error vary across elements and methods used. Although some harmonization errors will arise from the increased opportunity for processing errors, others will follow from modeling and imputation methods or the assumptions needed to align data for estimation. As with other sources of error for nonstatistical and integrated data, external gold standard data can be used to identify, reduce, and report harmonization error. Other approaches include benchmarking the harmonized data to known subject-specific and other relevant information and the use of sensitivity analysis. These approaches may identify harmonization errors that manifest for particular subsets or uses of the data, including estimates for subdomains. Data toolkits (Iwig *et al.* 2013, Seeskin *et al.* 2019) may facilitate harmonization efforts and reduce harmonization error through their provision of systematic information across sources for informing decisions and identifying appropriate sensitivity analysis.

Modeling Error: A special class of errors can be introduced when models are used to combine data. Models can be used to combine and integrate data at the record level or to combine estimates, possibly using covariates. To the degree that

assumptions that are associated with such models are inaccurate, they will lead to errors in the integrated data or estimates. Model-based methods for combining and integrating data include statistical matching or data fusion (Fosdick *et al.* 2016), imputation, Bayesian and/or hierarchical modeling, small area estimation, ensemble modeling, and calibration. New modeling approaches continue to be refined and developed, particularly for integrating emerging sources of data, such as satellite data or images. Machine learning and AI are special cases of modeling, albeit with a relatively greater reliance on empirical relationships in data on the outputs. Ensemble modeling is the process of fitting two or more related but different models and then combining the results in order to improve the predictive accuracy (Seni and Elder 2010, Singh *et al.* 2020). Ensemble methods in data analytics can add errors as outputs from initial models, which are subject to estimation error, are integrated into blended data, and used as inputs into other analytical methods (Singh, *et al.* 2020)

Modeling errors arise from inaccuracies in statistical model assumptions and from the effects of modeling decisions related to missing data, calibration variables and other constants, influential observations, and other factors. In addition, modeling error can differ among statistical outputs of the integrated data, including estimates produced for population domains and/or produced at various geographic and temporal resolutions.

Modeling error is often expressed using measures of statistical precision and statistical bias, such as standard errors and margins of error, which reflect sampling variability, and ‘goodness of fit’ statistics. A ‘best practice’ for statistical analysis continues to be segmenting data into training and testing sets to estimate these errors (Hastie *et al.* 2009). Modeling error for data products can also be evaluated using various approaches, including comparisons of model outputs to gold standard data, when available, and evaluations of the results using known subject-matter and other auxiliary information.

Sensitivity analysis (discussed in more detail below in section B.4) is commonly used to evaluate statistical models, particularly those that rely on unverifiable assumptions (Goldsmith 2015). New approaches to sensitivity analysis for model assessment continue to develop, particularly for machine learning and AI. For example, in a recent study, Lenis, *et al.* (2018) conduct a sensitivity analyses to investigate the consequences of model misspecification in the context of nonexperimental causal inference using a metric of model misspecification, termed the Degree of Misspecification, to explore which estimators are most sensitive to misspecification of either the treatment or the outcome model. In their example, the authors illustrate graphically through simulation how the degree of model misspecification for both the treatment and outcomes models increases the bias and the root mean square error of the effect estimate. Although developed in the context of nonexperimental causal inference, the approach can be extended when considering alternative models for integrating data. In another recent example, Siddique *et al.* (2019) use sensitivity analyses to address the influence of unverifiable assumptions involving the measurement error process. In particular, they address the assumptions of treatment and time invariance through the process by assessing the sensitivity of inferences on the effect of treatment when measurement error model parameters change from baseline to follow-up. Results of sensitivity analysis often show that alternative models are compatible with the data and known subject matter information but lead to different estimates.

The more complex the models or number of data inputs, the more decisions and data features could be assessed with sensitivity analysis. Although there is a need to move away from examining one parameter or aspect of the modeling process at a time to a more comprehensive approach, it is challenging to report aggregated numerical results across multiple dimensions. Bayesian and other Monte Carlo methods may be useful for assessing prior assumptions and multi-dimensional sensitivity. Directly incorporating the results of the sensitivity analyses into the final product can be done in some cases

and is an active research area. Bayesian Model Averaging has been used to incorporate the uncertainty about the model in the interval estimate of a statistic (Fragoso *et al.* 2018). Visualization tools that can display outcomes across multiple alternatives are promising (see *e.g.*, Rossen *et al.* 2020).

The Small Area Income and Poverty Estimates program at the Census Bureau funded by NCES, provides model-based estimates of income and poverty for the 3,141 counties and 13,206 school districts in the United States (Census Bureau 2020a). Estimates are produced by combining data from administrative records, postcensal population estimates, and the decennial census with direct estimates from the American Community Survey for single-year estimates. As modeling error can arise at various steps in the estimation process, the Census Bureau Quality Standards follows established standards for evaluating model-based estimates, including assessing model fit and performing sensitivity analysis (Census 2013).

In addition to documentation providing an overview of the source data, including its timeliness, and the estimation approach, standard errors for all Small Area Income and Poverty Estimates model-based estimates are produced as indications of their overall quality. The standard errors provided represent uncertainty due to the sampling variability within counties in the American Community Survey and “lack of fit” from the modeling process (Census 2017).

Additional Threats to Accuracy Involving Geographic Data: Many statistics are tied to individual locations and are easily displayed and analyzed with geographic information systems. The ease by which these systems allow the overlay of layers of data encourages spurious correlations among the layers and enables users to zoom in for geographic detail that is beyond the precision of the data.

Statistics about geographic areas are affected by how the areas are defined and subdivided. In addition to previously mentioned issues of spatial Granularity, the size of geographic areas significantly affects density measures. The population density of a city depends on how much surrounding rural area is included. How do the larger sizes of suburban census tracts compared to tracts in the center city affect neighborhood-level statistics? How are statistics on commercial activity in the census tracts affected by the delineation of census tracts to define residential areas along major streets that tend to divide commercial districts?

Distances between locations are often measured between centroids of the geographic units. Larger and fewer geographic units result in longer distances between units than smaller and more numerous geographic units for the same actual spatial distribution of activity. Even for small geographic units, is the activity being measured distributed evenly around the centroid, or is the activity being measured on the periphery of the geographic unit?

Changing the granularity for a data product does not always solve the statistical problems. Increasing the number of partitions to reduce the distortion of centroids can divide functionally similar areas and result in spatial autocorrelation and bias in linear models. Reducing the number of partitions can result in too much heterogeneity within the zones that overwhelms any analysis of differences between zones

A key challenge in geographic analysis is that neighborhood characteristics are often used as a proxy for individual characteristics because neighborhoods tend to attract similar residents and because a neighborhood can affect the character of its residents. These proxy measures add potential error from the variability among individual residents of the neighborhood as well as from effects that atypical residents can have on the behavior of others in the neighborhood.

B.4 Understanding and Minimizing Error

As discussed in the previous section, many of the sources of error in integrated data can be evaluated, conveyed, and minimized through the use of some combination of the three following techniques: comparison to gold standard, benchmarking to known information, and sensitivity analysis. The most appropriate technique or combination of techniques applied depends on the type of error of interest and the resources available, including expertise, time, funds, and relevant external data.

Comparison to Gold Standard: One method to evaluate threats for integrated data is to use a gold standard. A gold standard is often referred to as “an external source of truth” (Gordis 2014). When integrating sources through data linkage, Harron and colleagues describe the gold standard method as a way to “quantify errors (missed matches and false matches)” (Harron *et al.* 2017). However, finding a truth source/gold standard that is representative of the study at hand may be difficult to obtain.

Benchmarking to Known Information: A common practice to identify bias is to compare estimates to a benchmark or a standard. One example of this from integrating sources through data linkage is the assessment of linkage consent bias where estimates from the full population are compared to those who consented for their survey data to be linked. Sakshaug and Huber present this method as an assessment of absolute relative bias (Sakshaug and Huber 2015). A mitigation strategy if bias is found is to adjust survey sample weights for the sub-population that consented for linkage (Golden *et al.* 2015). However, there may be challenges with this method if there are systematic differences between the benchmark and the comparison group.

Sensitivity Analysis: Sensitivity analyses can be used to evaluate some threats to accuracy and reliability, particularly when using nonstatistical and integrated data. Although sensitivity analysis includes multiple methods and approaches, generally such an analysis involves varying unverifiable assumptions and decisions and examining their influences on the resulting outputs (Goldsmith 2005). Sensitivity analyses are particularly useful for evaluating errors unique or common for integrated data, such as modeling errors, harmonization errors, and linkage errors. As there are several approaches to integrating data, and resulting data products may have multiple outputs, sensitivity analyses take different forms when testing specific statistical assumptions, model specifications, linkage algorithms, influential observations, calibration variables and other constants, and other subject-specific assumptions and decisions. The results of sensitivity analyses can differ among data outputs from a common data file or data product, including estimates produced for population domains and/or at various geographic and temporal resolution. Although many analytic assumptions will be difficult to test or they will be untestable, the potential impact of violations of assumptions and benefits of potential mitigations are important in assessing accuracy and reliability. The more complex the integrated data product, the more data features that can be assessed with sensitivity analysis. Sensitivity analysis, however, can be time consuming and delay release of data or estimates.

B.13 Conclusion

Accuracy and reliability have been the cornerstone for assessing data quality. Data that are not accurate will not have high utility and will reduce the confidence users place in the data and the data producer. Identifying and measuring accuracy and reliability have been well-studied for survey data collected and disseminated by government agencies but are less developed for integrated data and for secondary uses of nonstatistical data, such as administrative records and data from satellites and other monitoring devices. As such, many of the recommendations in this report focus on research to improve the ability to identify and measure accuracy and reliability for new and emerging data products and new uses of data by federal

agencies. In addition to research for better understanding the threats described above, methods for measuring, evaluating and propagating the errors across all input sources and from the integration steps are needed. Further, methods for finding and using reference data, or gold standard data, to conduct these evaluations are needed. Disclosure risk protections affect accuracy and reliability and their impacts will increase with the complexity of data products and their ability to protect the data. Methods for finding the right balance between accuracy and reliability and protecting the confidentiality of the data continue to be needed.

References

- Blakely T, Salmond C. 2002. Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology* 31: 1246–1252.
- Bradley J, Wikle C, Holan S. 2016. Bayesian Spatial Change of Support for Count-Valued Survey Data. *Journal of the American Statistical Association*. 111(514): 472–487.
- Brillinger, David. 1986. The Natural Variability of Vital Rates and Associated Statistics. *Biometrics* 42, 693-734.
- Census Bureau. 2020. Small Area Income and Poverty Estimates (SAIPE) Program. Information available at <https://www.census.gov/programs-surveys/saipe.html>.
- Census Bureau. 2017. Quantifying Uncertainty in State and County Estimates. Available at <https://www.census.gov/programs-surveys/saipe/guidance/uncertainty.html>.
- Census Bureau. 2013. Statistical Quality Standards. Available at https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf.
- CNSTAT. National Academies of Sciences, Engineering, and Medicine, 2017. Innovations in federal statistics: Combining data sources while protecting privacy. National Academies Press.
- CNSTAT. National Academies of Sciences, Engineering, and Medicine, 2018. Federal statistics, multiple data sources, and privacy protection: Next steps. National Academies Press.
- Czajka JL, Stange M. 2018. Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines. Washington, DC: Mathematica Policy Research, April 27, 2018.
- Deming WE, Stephan FF. 1941. On the Interpretation of Census as Samples. *Journal of the American Statistical Association*. 36 (213): 45-49.
- Dusetzina SB, Tyree S, Meyer AM *et al.* 2014. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); An Overview of Record Linkage Methods. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK253312/>.
- European Statistical System. 2015. Quality Assurance Framework of the European Statistical System, Version 1.2., 2015. Available at <http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646>.
- Eurostat. 2020. European Statistical System Handbook for Quality and Metadata Reports, 2020 edition. 2020. Luxembourg: Publications Office of the European Union. Available at <https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf/bf98fd32-f17c-31e2-8c7f-ad41eca91783>.
- FCSM. 2001. Measuring and reporting sources of error in surveys. Washington, DC: U.S. Office of Management and Budget (Statistical Policy Working Paper 31).
- FCSM. 2020. A Systematic Review of Nonresponse Bias Studies in Federally Sponsored Surveys. Washington, DC. https://nces.ed.gov/fcsm/pdf/A_Systematic_Review_of_Nonresponse_Bias_Studies_Federally_Sponsored_SurveysFCSM_20_02_032920.pdf.

Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys. 2016. Current measures of sexual orientation and gender identity in federal surveys.

Fellegi IP, Sunter AB. 1969. A theory for record linkage. *Journal of the American Statistical Association*. 64(328):1183-210.

Fosdick BK, DeYoreo M, Reiter JP. 2016. Categorical data fusion using auxiliary information. *Annals of Applied Statistics*. 10(4):1907-1929. doi:10.1214/16-AOAS925. <https://projecteuclid.org/euclid.aoas/1483606845>.

Fragoso TM, Bertoli W, Louzada F. 2018. Bayesian Model Averaging: A Systematic Review and Conceptual Classification. *International Statistical Review*. 86:1– 28. doi: [10.1111/insr.12243](https://doi.org/10.1111/insr.12243).

Golden C, Driscoll AK, Simon AE, et al. Linkage of NCHS population health surveys to administrative records from Social Security Administration and Centers for Medicare & Medicaid Services. National Center for Health Statistics. *Vital Health Stat* 1(58). 2015.

Goldsmith CH. 2005. Sensitivity Analysis. *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd. DOI: 10.1002/0470011815.b2a04051.

Gordis, L. (2014). *Epidemiology* (Fifth edition.). Philadelphia, PA: Elsevier Saunders.

Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc AMIA Symp*. 2002;305-309.

Groves R, Fowler F, Couper M, Lepkowski J, Singer E, Tourangeau R. 2009. *Survey Methodology*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons.

Groves R. 1989. *Survey Errors and Survey Costs*. Hoboken, New Jersey: John Wiley & Sons.

Handwerker EW, Kim MM, Mason LG. 2011. Domestic employment in U.S.-based multinational companies. *Monthly Labor Review*. 134 (10).

Harron K, Hagger-Johnson G, Gilbert R, Goldstein H. 2017. Utilising identifier error variation in linkage of large administrative data sources. *BMC Medical Research and Methodology*. 17(1):23. Published 2017 Feb 7. doi:10.1186/s12874-017-0306-8.

Harron, KL, Doidge, JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, van der Meulen JH. 2017. A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology*, Volume 46, Issue 5, October 2017, Pages 1699–1710, <https://doi.org/10.1093/ije/dyx177>.

Hastie T, Tibshirani R, Friedman J. 2009. *Elements of Statistical Learning*. Second edition. Springer Series in Statistics.

Hof MH, Anita C, Ravelli AC, Zwinderman AH. 2017. A Probabilistic Record Linkage Model for Survival Data. *Journal of the American Statistical Association*. 112:520, 1504-1515, DOI: 10.1080/01621459.2017.1311262.

Holan S. 2018. Recent Advances in Spatial and Spatio-Temporal Change of Support for Official Statistics. Presented to the FCSM/Washington Statistical Society Workshop on Quality of Integrated Data. Reporting on Quality Issues in Data Processing, January 25, 2018. www.washstat.org/presentations.

ICSP. 2018. Principles for Modernizing Production of Federal Statistics, Available at <https://nces.ed.gov/fcsm/pdf/Principles.pdf>.

Iwig W, Berning M, Marck P, Prell M. 2013. Data quality assessment tool for administrative data. Prepared for a subcommittee of the FCSM, Washington, DC. <https://nces.ed.gov/FCSM/pdf/DataQualityAssessmentTool.pdf>.

Jang, D. 2018. Data Harmonization in Survey Data Integration, presented to the FCSM/Washington Statistical Society Workshop on Quality of Integrated Data. Reporting on Quality Issues in Data Processing, January 25, 2018. https://nces.ed.gov/FCSM/pdf/20180125_Jang.pdf.

Kish L, Frankel MR. 1970. Balanced Repeated Replications for Standard Errors. *Journal of the American Statistical Association*. 65 (331): 1071–1094. JSTOR, www.jstor.org/stable/2284276.

Krewski D, Dewanji A, Wang Y *et al.* 2005. The effect of record linkage errors on risk estimates in cohort mortality studies. *Survey Methodology* 31:13–21.

Lahiri P, Larsen MD. 2005. Regression Analysis with Linked Data, *Journal of the American Statistical Association*. 100:222-230, doi: [10.1198/016214504000001277](https://doi.org/10.1198/016214504000001277).

Lenis D, Ackerman B, Stuart EA. 2018. Measuring model misspecification: Application to propensity score methods with complex survey data. *Computational Statistics & Data Analysis*. 128: 48-57. <https://www.sciencedirect.com/science/article/pii/S0167947318301105>.

Massey CG, O'Hara A. 2014. Person Matching in Historical Files using the Census Bureau's Person Validation System (No. 2014-11). Center for Economic Studies, U.S. Census Bureau. Available at <https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-11.html>.

NCHS. Collaborative Center for Questionnaire Design and Evaluation Research. 2020. Q-bank. Information found at <https://wwwn.cdc.gov/qbank/Home.aspx>. Accessed May 1, 2020.

NCHS. Division of Analysis and Epidemiology. 2019. The Linkage of the 2016 National Hospital Care Survey to the 2016/2017 National Death Index: Methodology Overview and Analytic Considerations, August 2019. Hyattsville, Maryland. Available at the following address: <https://www.cdc.gov/nchs/data-linkage/index.htm>.

Newcombe HB, Kennedy JM, Axford SJ, James AP. 1959. Automatic linkage of vital records. *Science*. 130(3381):954-959. doi:10.1126/science.130.3381.954.

OMB. 2006. Standards and Guidelines for Statistical Surveys. September 2006. Available at: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

OMB. 2008. Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies (73 FR 12621, March 7, 2008).

OMB. 2014. Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units (79 FR 71609, Dec 2, 2014).

OMB. 2019. M-19-15: Improving Implementation of the Information Quality Act. April 2019. Available at: <https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf>

Resnick D, Asher J. 2019. Measurement of Type I and Type II Record Linkage Error. In Proceedings of the Joint Statistical Meetings, Government Statistics Section. Alexandria, VA: American Statistical Association.

Rossen LM, Womack LS, Hoyert DL, Anderson RN, Uddin SFG. 2020. The Impact of the Pregnancy Checkbox and Misclassification on Maternal Mortality Trends in the US, 1999-2017. National Center for Health Statistics. *Vital Health Stat* 3(44). Available at <https://www.cdc.gov/nchs/maternal-mortality/dashboard/index.htm>

- Sayers J, Campbell S, Thompson C, Jackson G. 2018. Data Linkage with an Establishment Survey. In Proceedings of the Federal Committee on Statistical Methodology Research Conference, Session G-4.
- Schenker N, Gentleman JF, Rose D, Hing E, Shimizu IM. 2002. Combining estimates from complementary surveys: a case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*. 117:393–407.
- Seeskin ZH, Ugarte G, Datta AR. 2019. “Constructing a toolkit to evaluate quality of state and local administrative data.” *International Journal of Population Data Science* 4.1.
- Seni G, Elder JF. 2010. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan and Claypool.
- Sakshaug, J.W. and M. Huber, An Evaluation of Panel Nonresponse and Linkage Consent Bias in a Survey of Employees in Germany. *Journal of Survey Statistics and Methodology*, 2015. 4(1): p. 71-93.
- Siddique J, Daniels M, Carroll R, Raghunathan TE, Stuart E, Freedman L. 2019. [Measurement error correction and sensitivity analysis in longitudinal dietary intervention studies using an external validation study](#). Presented at the FCSM/WSS Workshop on Sensitivity Analysis with Integrated Data. June 10, 2019. Available at: <http://www.washstat.org/presentations/>.
- Singh L, Traugott M, Bode L, Budak C, Davis-Kean PE *et al.* 2020. Data Blending: Haven't We Been Doing This for Years? White paper from the Massive Data Initiative. Georgetown University. Found at https://www.jonathanmladd.com/uploads/5/3/6/6/5366295/mdi_data_blending_white_paper_-_april2020.pdf.
- Wine J, Siegel P, Stollberg R. 2018. 2015-16 National Postsecondary Student Aid Study (NPSAS:16) Data File Documentation (NCES 2018-482). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved May 13, 2020 from <http://nces.ed.gov/pubsearch>.
- Zhang, LC. 2012. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41-63.