Transparent Reporting for Integrated Data Quality: Practices of Seven Federal Statistical Agencies

FCSM-19-01

Prepared by

Mark Prell (Co-Chair), Chris Chapman (Co-Chair), Samson Adeshiyan, Dennis Fixler, Tom Garin, Lisa Mirel, and Polly Phipps

under the auspices of
Federal Committee on Statistical Methodology

Statistical and Science Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

September 2019

Recommended citation: Prell, Mark, Chris Chapman, Samson Adeshiyan, Dennis Fixler, Tom Garin, Lisa Mirel, and Polly Phipps. 2019. Transparent Reporting for Integrated Data Quality: Practices of Several Federal Statistical Agencies. FCSM 19-01. Federal Committee on Statistical Methodology. September 2019.

THE FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

MEMBERS

John Eltinge (Co-Chair)
U.S. Census Bureau

Rochelle (Shelly) Martinez (Co-Chair)
Office of Management and Budget

Nancy Bates
U.S. Census Bureau

Chris Chapman
National Center for Education Statistics

Jeffrey Chen
Bureau of Economic Analysis

Jennifer Edgar
Bureau of Labor Statistics

Dennis Fixler
Bureau of Economic Analysis

Barry Graubard
National Cancer Institute

Jennifer Madans
National Center for Health Statistics

Wendy Martinez
Bureau of Labor Statistics

Jaki McCarthy
National Agricultural Statistics Service

Jennifer Nielsen (Secretary)
National Center for Education Statistics

Jennifer Parker
National Center for Health Statistics

Polly Phipps
Bureau of Labor Statistics

Nancy Potok (Champion)
Statistical and Science Policy
Office of Management and Budget

Mark Prell
Economic Research Service

Joseph Schafer
U.S. Census Bureau

Rolf Schmitt
Bureau of Transportation Statistics

Marilyn Seastrom
National Center for Education Statistics

Joy Sharp
Energy Information Administration

Bruce Taylor (Webmaster)
National Center for Education Statistics

G. David Williamson
Centers for Disease Control and Prevention

Linda Young
National Agricultural Statistics Service

MEMBERS OF THE TRANSPARENT REPORTING PROJECT

Mark Prell (Co-Chair)
Economic Research Service

Chris Chapman (Co-Chair)
National Center for Education Statistics

Samson Adeshiyan
National Center for Science and Engineering Statistics

Dennis Fixler
Bureau of Economic Analysis

Tom Garin
National Center for Veterans Analysis and Statistics

Lisa Mirel
National Center for Health Statistics

Polly Phipps
Bureau of Labor Statistics

Transparent Reporting for Integrated Data Quality: Practices of Seven Federal Statistical Agencies

Contents

Executive Summary vi

Acknowledgements xiv

List of Acronyms xv

1. Introduction 1
2. National Accounts 24
3. Integration of Administrative Data and Survey Data 34
4. Proprietary Data 57
5. Integration of Data from Multiple Surveys 74
6. User Assessments of Transparent Reporting 83
7. Challenges for Improving Transparent Reporting 93
8. Conclusion 97

Appendix 1. Definitions of Data Quality Dimensions 98

Appendix 2. Instrument for Users Identified as Informed Consumers 100

Appendix 3. Instrument for Users Identified as Researchers 106

Appendix 4. Item Responses to Customer Survey, Full Sample and by Type of User 112

Appendix 5. Means of Item Responses to Customer Survey, Full Sample and by Type of User and Statistical Product 116

Appendix 6. Detailed Statistical Results of the Customer Survey 121

Appendix 7. Alternative Data Qualitative Assessment Scorecard (Bureau of Labor Statistics) 124

Appendix 8. Alternative Data Quantitative Assessment Guide (Bureau of Labor Statistics) 127

Executive Summary

Goals. This report provides results from the Transparent Reporting Project, which examined how seven agencies in the federal statistical system assess the quality of integrated data used in one of the agency's statistical products and how they report data quality to users of the product. The project was designed to serve the information needs of the project's participating agencies, the members of the Interagency Council on Statistical Policy (ICSP), and agencies throughout the federal statistical system. In this report, the terms "statistical agency" or "federal statistical agency" can refer to any agency or unit in the federal statistical system, not necessarily one of the 13 principal statistical agencies.

Transparent reporting is achieved when an agency provides clear and detailed documentation so that users can assess for themselves the quality of the data. As seen from a user's perspective, data quality refers to the data's "fitness for use" for a user's own needs. A statistical agency has a responsibility to convey information to users about data quality. In addition, transparent reporting can be a step or stage in a process of continuous improvement by which the statistical agency enhances the quality of the data.

Integrated data are also known as multiple-source, blended, hybrid, or mixed data. Examples of integrated data include the country's national accounts or a dataset that results from linking survey and administrative data. As new data sources and methodologies develop, federal statistical agencies are using integrated data more extensively than in the past. Less is known about the quality of integrated data, and their non-survey sources of data (such as administrative data or proprietary data), than about the quality of survey data.

The Transparent Reporting Project had four specific goals:

Goal 1. examine current practices (among the agencies participating in the project) for transparent reporting about statistical products based on integrated data;

Goal 2. assemble available estimates, in dollars or staff time, for costs associated with reporting on integrated data quality;

Goal 3. solicit user evaluations of the quality of integrated data quality reporting; and

Goal 4. identify possible improvements to existing practices that could result in more transparent reporting to better inform users about integrated data quality.

The scope of the Transparent Reporting Project was limited to statistical products of the seven participating agencies. An examination of these products and their documentation can provide examples of current practices and potential improvements for other statistical agencies.

The project was conducted by a cross-agency team under the auspices of the Federal Committee on Statistical Methodology (FCSM) Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources.

Seven statistical products. For its case studies, the Transparent Reporting Project examined seven statistical products, classified using four domains of interest. The first domain was *national accounts statistics*, for which the project team examined the Personal Consumption Expenditures component of Gross Domestic Product (Bureau of Economic Analysis (BEA)). The second domain was *integrating administrative and survey data*, which included three statistics products: the National Postsecondary Student Aid Study (National Center for Education Statistics (NCES)); linkage between the National Hospital Care Survey and the National Death Index (National Center for Health Statistics (NCHS)); and the 2016 Veteran Population Model (National Center for Veterans Analysis and Statistics (NCVAS)). The third domain was the *use of proprietary data, possibly in combination with other data*, for which the project examined the telecommunications component of the Consumer Price Index (CPI) (Bureau of Labor Statistics (BLS)) and the National Household Food Acquisition and Purchase Study (FoodAPS) and other uses of proprietary data (Economic Research Service (ERS)). The fourth domain was *integration of data from multiple surveys*, for which the project team examined the Scientists and Engineers Statistical Data System (National Center for Science and Engineering Statistics (NCSES)).

Methodology. The Transparent Reporting Project drew upon three sources of information. First, the project team reviewed documentation by which agencies assess data quality and report on it to users (Goal 1). Second, the team developed a web-based customer survey that solicited user appraisals of the quality of the data, user satisfaction with the transparency of the agency reporting on data quality, and user suggestions for improving data quality and documentation (Goal 1, 3, and 4). Third, agency staff were queried for rough estimates of the costs in terms of dollars or staff time to prepare and disseminate agency documentation on data quality (Goal 2).

To examine how documentation describes data quality, the project used a data quality framework that has been used at the BEA and elsewhere. The framework consists of eight dimensions of data quality: *relevance*, *accuracy*, *reliability*, *timeliness*, *punctuality*, *consistency*, *comparability*, and *access*. This set of dimensions reflects the influence of various data quality frameworks, including those laid out by the Organization for Economic Cooperation and Development, which in turn drew from the quality assurance framework of the European Statistical System. While these eight dimensions of data quality originated in the context of survey data, they have been extended by U.S. federal statistical agencies to administrative data and integrated data. These dimensions have been discussed in the 2017 report *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* written by an expert panel convened by the Committee on National Statistics of the National Academies of Science, Engineering, and Medicine.

The Transparent Reporting Project's customer survey represents a pioneering effort—the first known survey designed to collect assessments of transparent reporting on integrated data using common items across users of several statistical products produced by different agencies. The survey was based on a purposive sample of selected users rather than a probability-based sample of all users of the agencies' statistical products. In total across the project agencies, 105 selected users were contacted to participate in the survey.

A second feature of the project's customer survey is that it developed two instruments for the customer survey, tailoring an instrument for each of two types of users. The project used the term "informed consumers" for those who use results based on the integrated data product and need information on how the product was developed to interpret the information it provides. The term "researchers" was used for those who either directly use raw data in the integrated data product or need detailed technical information about the product for their research activities. The project's two instruments shared many items in common, while other items appeared on only one of the two instruments. Altogether, there were 68 distinct items: 47 common items that appeared on both instruments, 6 items posed only to informed consumers, 12 items posed only to researchers, and 3 items (that appeared on one or both instruments) that were posed only for certain agencies' users.

The survey was designed to provide a rich array of information on many aspects of documentation and, to a lesser extent, on data quality. Some items can be grouped into four clusters that did not address documentation: Preliminary (2 items); Dimensions of Data Quality (8 items); Overall Assessment of confidence in the data or statistical product (1 item); and Improve Data Quality (1 item). Six clusters that focused on documentation were: Quality of Documentation (8 items); Source Data (8 items); Agency Evaluation of Quality (5 items); Data Integration (10 items); How to Use the Product (14 items); Agency Contacts (8 items). The final cluster—Agency Contacts—shows that the project conceived of "documentation" broadly, including both materials written for users and user communication with agency staff through telephone, e-mail, etc. The customer survey posed questions about "agency contacts" because some users contact agency staff for information similar to what may be provided in written documentation, and staff may themselves turn to that documentation for addressing questions.

Users rated some items based on a familiar Likert scale (Very dissatisfied (=1); Somewhat dissatisfied (=2); Neither dissatisfied nor satisfied (=3); Somewhat satisfied (=4); and Very satisfied (=5)). Most other items were dichotomous or allowed for unstructured, free-text responses.

Key Findings. The eight dimensions of data quality adopted for the Transparent Reporting Project (relevance, accuracy, reliability, timeliness, punctuality, consistency, comparability, and access) are compatible with, and implications of, a set of core concepts (quality, objectivity, utility, and integrity) in information quality guidelines established by the Office of Management and Budget and departments across federal government. The project found that agency documentation provides information on these dimensions. At the same time, the statistical products under study were quite diverse, combining data from multiple sources using various methodologies. As a result, agencies can be providing to users the type of information to be expected, based on information quality guidelines, but reporting it in different ways. The project's data quality framework and the documentation practices reviewed by the project can serve as illustrations of how to convey information on integrated data quality to their users.

Findings from the customer survey are based on (unweighted) responses from users who choose to participate, with no adjustment made for non-response. Of the 105 users who were asked to participate in the survey, 46 participated for an overall response rate of 43.8 percent (46/105).

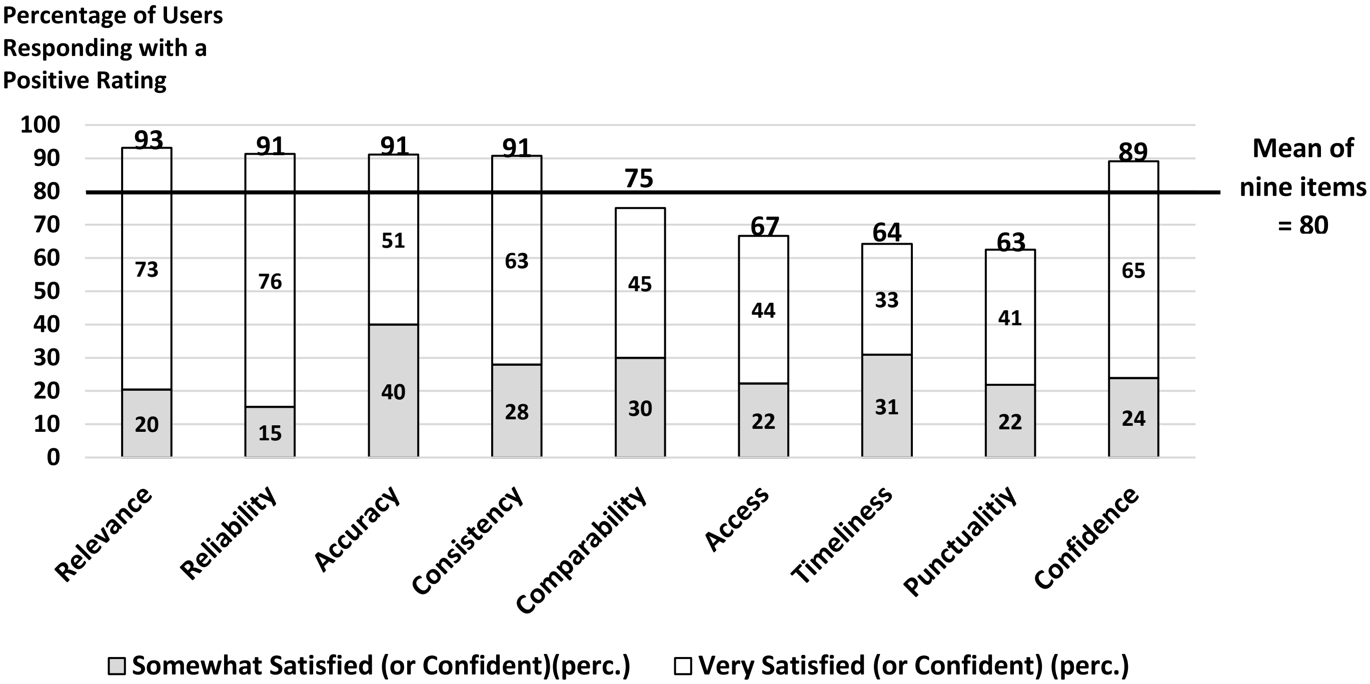
There was a notable difference in response rates between the two subgroups of users. Only 8 of the 31 informed consumers contacted for the survey chose to participate, representing a response rate of about 25.8 percent. Researchers participated at a rate of 51.3 percent, with 38 researchers participating out of 74 contacted. The difference between the subgroups' response rates was just over 25 percentage points; alternatively, the response rate of the researchers was about double the rate of the informed consumers.

Because the sample was small and purposive, results cannot be generalized to all users of the agencies' statistical products. Instead, results from the customer survey are interpreted as direct measures of the responses of the participating users and as illustrative or suggestive of how other, non-sampled users may assess data quality and documentation.

The customer survey provided results that identify, from a user perspective, both areas of strength and components of data and data quality documentation that are not rated as highly by users. While results indicate that in general users are satisfied with both the quality of data and the quality of documentation, there is still room for improvements.

For each of the eight dimensions of data quality, figure 1 shows the percentage of respondents who provided a positive rating of either Somewhat Satisfied or Very Satisfied (4 or 5 on the Lickert scale). A single item was used to assess each dimension.

Figure 1. User Assessments for Dimensions of Data Quality and User Overall Confidence in the Data



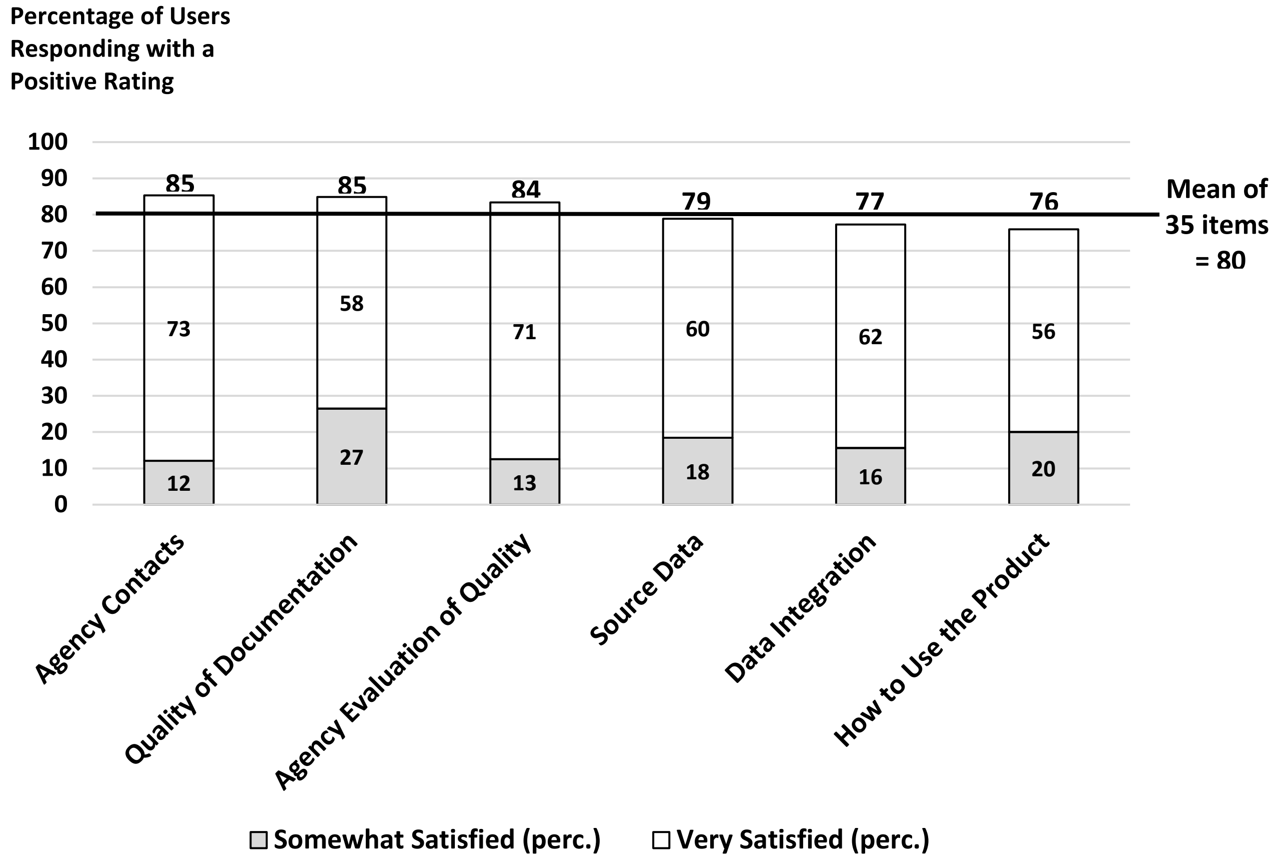
Note: A positive rating is a response of "Somewhat satisfied" or "Very Satisfied" on a Likert scale assessing each dimension of data quality and "Somewhat confident" or "Very confident" when describing overall confidence in the data. A percentage of positive ratings is displayed above each box.
Source: Transparent Reporting Project Customer Survey.

A cluster of four dimensions that were assessed relatively highly, with over 90 percent of users providing a positive rating, were relevance (93 percent), accuracy (91 percent), reliability (91 percent) and consistency (91 percent). Dimensions for which the percentage of positive ratings was relatively low were timeliness (64 percent), punctuality (63 percent), comparability (75 percent), and access (67 percent).

A bottom-line assessment of data quality is captured by a user assessment of confidence in the data. The item asked "Overall, how confident are you in the data or the statistics that you obtain from {Product}?" on a Likert scale ranging from "Very unconfident" up to positive ratings at "Somewhat confident" and "Very confident." The figure shows the percentage of positive ratings for overall confidence to be relatively high at 89 percent. The arithmetic mean for the item on overall assessment of confidence and for the cluster of eight items on Data Quality is about 80 percent; as an unweighted mean, that figure does not take consider that different numbers of users responded to the nine items.

In general, users were satisfied with documentation about data as shown in figure 2. For the 35 Likert items in the six clusters of related items on documentation, the items were averaged separately, by cluster, to obtain intra-cluster (arithmetic) means.

Figure 2. User Assessments of Agency Documentation, by Cluster of Related Items



Note: A positive rating is a response of "Somewhat satisfied" or "Very satisfied" on a 5-point Likert scale. A mean percentage of positive ratings for a cluster of related items is displayed above each box. The mean percentages of "Somewhat satisfied" and "Very satisfied" may not sum exactly to the mean percentage of positive ratings for a cluster.
Source: Transparent Reporting Project Customer Survey

From a user's perspective, areas of relative strength in the transparent reporting on data quality were Agency Contacts (obtaining information from agency staff) and general-level Quality of Documentation, each of which had means of about 85 percent. The means for Agency Evaluation of Quality and for Source Data were 84 and 79 percent, respectively. Two clusters for which the percentages of users with positive ratings were relatively low were Data Integration and How to Use the Product, each with means of 77 and 76 percent. Averaging across the 35 items on documentation, mean of the percentages of users with positive ratings was 80 percent.

Aggregating results from 35 individual items to obtain intra-cluster means provided useful summary statistics for figure 2. However, the process of averaging also blends together relatively high or low item assessments into the composite averages. Examining user assessments at the item level rather than the cluster level provides additional detail on strengths and areas for potential improvement.

At the level of individual items, relatively high percentages of users expressed positive ratings of satisfaction for such items as: "How representative information from {Product} is of the population or concept you are using it to study (item #47)"—100 percent; "The information provided [by agency staff] effectively addressed questions about {Product} (#58)"—97 percent; "Detail with which the {Agency} explained the purpose for which the source information was collected initially (#24)—89 percent; and "Detail [from agency staff] about how the sources of information were integrated into {Product} (#62)"—86 percent; item numbers refer to a master listing of customer survey items provided in Appendices 4 and 5.

Items with relatively low percentages include: "[Details on] [r]eferences to find out more about the source data beyond what was included in the {Agency}'s documentation (#25)"—67 percent; "How the product was adjusted to prevent disclosure of respondent or subject identity (#55)"—64 percent; and "How errors from each data source affect the overall error for the product (#53)"—58 percent.

In the areas where users express relatively less satisfaction, the difficulty or ease with which to improve documentation varies across items and across statistical products for a given item. For example, documentation could potentially provide more references about the source data (#25). Adding additional references may be straightforward for some statistical products. However, in some instances such references may not be available, perhaps especially when data are proprietary. Perhaps more detail can be added to documentation to explain steps the agency took to prevent disclosure of confidential data (#55). At the same time, a public explanation of such methods can be incomplete because revealing too much information defeats the purpose of the methods; a short note explaining the need to limit public information may satisfy users who otherwise might expect more detail. It is inherently difficult with multiple-source data to identify how errors in data sources affect the overall error for the product (#53). Moreover, error in multiple-source data needs to be evaluated in relation to the use of the data. This topic is on the frontier of research in integrated data, representing an extension of classic questions concerning Total Survey Error and how various types of sampling and non-sampling errors affect overall error in survey-based data or statistical product. Even before additional research is completed, though, agencies can strive to be transparent in discussing source data, errors in the source data, and methods by which data were integrated for the statistical product.

User assessments are not the only information an agency might use to plan for improvements in documentation. Various opportunities for improvement have cost implications for agency resources. The Transparent Reporting Project found that cost estimates for providing current levels of documentation varied widely across the agencies and their statistical products and programs. The one-time costs to prepare materials for users for one of the statistical products may be in the range of about $15,000 to $90,000, while the cost of documentation of all BEA programs is about $2 million annually. In addition, the project found that such costs estimates are difficult to develop—agencies do not structure budgets to include documentation as a line item. The project also found that the resulting cost estimates are difficult to compare across statistical products because the estimates can include some costs besides the documentation costs of the statistical product under study.

An agency can use its budget to improve documentation or to improve the quantity or quality of data. Improvements in documentation entail costs that need to be weighed against how users would value other applications of an agency budget.

Finally, to reiterate, we recognize that the small non-probability sample and number of responses limit the strength of any inferences. However, we believe that the results can serve as useful indicators for agencies to examine their current practices.

Acknowledgements

The report is based in part on responses to a customer survey that asked survey participants to assess a specified statistical product and its documentation. While the individuals who participated in the customer survey are not named here, the authors acknowledge the importance of the survey participations and express gratitude for their contribution.

The report's authors also thank other people whose expertise and assistance contributed to the report:

Bureau of Labor Statistics
Craig Brown
Crystal Konny

Economic Research Service
Andrea Carlson
Mark Denbaly
David Levin (currently at Food Safety and Inspection Service)
Megan Sweitzer

General Services Administration
Trey Bradley

Grant Thornton, LLP
Christiana Chen
Tracy Jones
Sophia Minassian
Esther O'Mara

RTI International
Johnathan Conzelmann
Austin Lacy

List of Acronyms

BEA	Bureau of Economic Analysis
BLS	Bureau of Labor Statistics
CNSTAT	Committee on National Statistics
ERS	Economic Research Service
ESS	European Statistical System
FCSM	Federal Committee on Statistical Methodology
FoodAPS	National Household Food Acquisition and Purchase Study
FNS	Food and Nutrition Service
FSS	Federal Statistical System
GDP	Gross Domestic Product
GSA	General Services Administration
GT	Grant Thornton
ICSP	Interagency Council on Statistical Policy
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
NDI	National Death Index
NHCS	National Hospital Care Survey
NPSAS	National Postsecondary Student Aid Study
OECD	Organization for Economic Cooperation and Development
OMB	Office of Management and Budget
SESTAT	Scientists and Engineers Statistical Data System
SNAP	Supplemental Nutrition Assistance Program
Stats NZ	Statistics New Zealand
TSE	Total Survey Error
UN	United Nations
US	United States
USDA	U.S. Department of Agriculture
VA	U.S. Department of Veterans Affairs
VetPop 2016	Veteran Population Projection Model 2016
WIC	Special Supplemental Nutrition Program for Women, Infants, and Children

Transparent Reporting for Integrated Data Quality:
Practices of Seven Federal Statistical Agencies

1. Introduction.

Background. In 2017, the Federal Committee on Statistical Methodology (FCSM) established its Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources (hereafter, FCSM Working Group). The charge of the FCSM Working Group is "to identify best practices for reporting on quality dimensions of integrated data products created and produced by federal statistical agencies" (Federal Committee on Statistical Methodology Working Group, 2018, p. 3); in this report, the terms "statistical agency" or "federal statistical agency" can refer to any agency or unit in the federal statistical system, not necessarily one of the 13 principal statistical agencies. This report, Transparent Reporting for Integrated Data Quality: Practices of Seven Federal Statistical Agencies, contributes to that goal and is the FCSM Working Group's second report. The FCSM Working Group's first report, Transparent Quality Reporting in the Integration of Multiple Data Sources: A Progress Report, 2017-2018, summarizes the activities and findings of the Working Group during its inaugural year.

FCSM established the FCSM Working Group in response to a request by the Interagency Council on Statistical Policy (ICSP) for detailed analysis on the statistical quality of integrated data. At the time of the request, the ICSP included the heads of the principal federal statistics agencies, representatives of other statistical agencies under rotating membership (the National Center for Veterans Analysis and Statistics), and the Chief Statistician of the United States from the Office of Management and Budget.

Motivation and Structure for the Transparent Reporting Project. In May 2018, the ICSP requested that the FCSM conduct a study on practices used by ICSP agencies to report in a clear and transparent fashion on the quality of integrated data. Transparent reporting of data quality is achieved when an agency provides clear and detailed documentation so that users can assess data quality—that is, the data's "fitness for use"—for themselves. It can be a natural process for quality improvement to begin with assessing, measuring (where possible), and reporting on current features of data quality. Transparent reporting can be a first step in continuous improvement by which the statistical agency enhances the quality of the data.

The Transparent Reporting Project had four goals:

- examine current practices (among the agencies participating in the project) for transparent reporting about statistical products based on integrated data; integrated data is also known as blended, hybrid, or mixed data such as a dataset that results from linking survey and administrative data
- assemble available estimates, in dollars or staff time, for costs associated with reporting on integrated data quality;
- solicit user evaluations of the quality of integrated data quality reporting; and
- identify possible improvements to existing practices that could result in more transparent reporting to better inform users about integrated data quality.

The broader vision behind these goals was that the project's agencies, the members of the ICSP, and agencies throughout the federal statistical system could learn about transparent reporting on integrated data products, and ways to improve it, from the practices of the project's seven agencies. In addition, transparent reporting can be a critical step in the process by which a statistical agency enhances the quality of the data. That is, in preparation for improving data quality, an agency can begin by assessing and (where possible) measuring its current level of quality across different dimensions and by providing documentation in sufficient detail so that the user can determine fitness-for-use of currently available data.

In the time since the ICSP asked FCSM to conduct the study, the ICSP issued its Principles for Modernizing Production of Federal Statistics (Interagency Council on Statistical Policy, 2018). The ICSP considers integrated data to be a key element of initiatives to strengthen the Federal Statistical System. One of the principles is that, as they serve their missions, agencies should use sources of information that are both the "highest quality" and "reasonably attainable," including "non-statistical data sets and derivative information" (Interagency Council on Statistical Policy, 2018, p. 3). Another principle is that "Agencies should report transparently on the quality of information they disseminate." The work of the ICSP and the FCSM on integrated data fits under the Federal Data Strategy for leveraging data as a strategic asset, which is part of the 2018 President's Management Agenda (Interagency Council on Statistical Policy, 2018, p. 2).

In response to the ICSP request, the FCSM solicited agency participation for the study. Beginning in June 2018, seven statistical agencies on the ICSP began collaboration on the Transparent Reporting Project. The agencies selected statistical products, each based on integrated data, for review and assessment in seven case studies.

For the case studies, members of the project team examined how documentation of each agency's statistical product described data quality for its integrated data product. In addition, the agencies conducted a customer survey that used common, cross-agency items on transparent reporting on integrated data. Using newly developed instruments, the survey solicited from the agencies' users their assessments of how transparently—that is, how clearly and completely—the agency reports on the data quality of integrated data products so that users can determine if the product can be used for their purposes. The survey also identified, from user perspective, what gaps there may be in documentation and what improvements may be desired in data quality of the statistical product itself. The customer survey was intended to be exploratory and illustrative—a first attempt at soliciting user responses on this complex subject.

The structure of the Transparent Reporting Project was designed to examine each of four broad categories of integrated data identified by the ICSP. The four categories, together with the agencies' seven case studies for the project, depicted in Table 1.1.

Table 1.1. Seven Statistical Products in the Transparent Reporting Project's Case Studies

Statistical Products, by categories of data	Agency
I. National accounts statistics	
(1) Personal Consumption Expenditures (PCE) component of Gross Domestic Product	Bureau of Economic Analysis (BEA), U.S. Department of Commerce
II. Integration of administrative and survey data	
(2) National Postsecondary Student Aid Study (NPSAS)	National Center for Education Statistics (NCES), U.S. Department of Education
(3) Linkage between the National Hospital Care Survey and the National Death Index (NHCS-NDI Linkage)	National Center for Health Statistics (NCHS), U.S. Department of Health and Human
(4) Veteran Population Project Model 2016 (VetPop 2016)	National Center for Veterans Analysis and Statistics (NCVAS), U.S. Department of Veterans Affairs.
III. Proprietary data, possibly in combination with other data	
(5) Telecommunications component of Consumer Price Index (CPI);	Bureau of Labor Statistics (BLS), U.S. Department of Labor
(6) National Household Food Acquisition and Purchase Survey (FoodAPS) and other uses of proprietary data	Economic Research Service (ERS), U.S. Department of Agriculture
IV. Integration of data from multiple surveys	
(7) Scientists and Engineers Statistical Data System (SESTAT)	National Center for Science and Engineering Statistics (NCSES), National Science Foundation

The project also assembled information from each of the seven agencies about the cost, in terms of staff time or dollars, for producing the agency's reports on its integrated data product. This information helps place their documentation efforts in context and provides some basis for assessing the costs associated with different scales and types of reporting efforts. Cost estimates were difficult to obtain because agency budgets and accounting frameworks are not structured to distinguish the cost of reporting as a line item separate from other costs.

The project found that estimated costs of reporting on data quality vary widely. In part, that variation reflects differences in the purpose, size and scope of the products and their associated documentation, which is the variation the project sought to measure. In addition, the variation reflected other costs that were mixed in, including costs of reporting on other statistical products (besides the one in question) or costs of preparation of the statistical product (which can be difficult to separate from documentation). Costs estimates are provided in the separate chapters as each product is discussed. Costs estimates should all be considered highly approximate.

OMB memo M-19-15, Improving Implementation of the Information Quality Act, provides an update to the Information Quality Act to "reinforce, clarify, and interpret agency responsibilities" (Office of Management and Budget, 2019, p. 1). While M-19-15 was issued after the Transparency Reporting Project was designed and its customer survey fielded, the findings and agency practices described in this report can help agencies as they implement the Information Quality Act in light of M-19-15. For example, the memo's Implementation Update 2.2 states, in part, that "Agencies should provide the public with sufficient documentation about each dataset released to allow data users to determine the fitness of the data for the purpose for which third parties may consider using it" (Office of Management and Budget, 2019, p. 5). Statistical agencies are different in terms of how they produce a variety of statistical products based on integrated data. They are similar in that they need to document and communicate to users the types of data and the data's quality characteristics so that users can determine fitness for use.

Types of Data. Federal statistical agencies and programs provide data to inform decisions of the public, businesses and government (Office of Management and Budget, 2018, pp. 3-6). The highly decentralized federal statistical system comprises over 100 agencies (the formal names of which may use the terms center, institute, bureau, etc.). The agencies of the federal statistical system develop and disseminate many types of statistical products including datasets containing micro-level observations, estimates (single-number statistics such as totals, means, or regression coefficients) and statistical reports. For evidence-based decision-making, these statistical products need to be relevant, accurate and timely.

Currently, most federal statistical products are based on censuses and probability surveys (Jarmin, 2019, pp. 165-167; National Academies of Sciences, Engineering, and Medicine, 2017a, pp. 21-29); a probability survey is based on each element of a population or universe having a known probability of selection into the sample, in contrast to a convenience sample or a purposive survey. However, several factors are making reliance on traditional, survey-based methods of federal data collection problematic. Response rates to federal surveys are declining. Relatedly, the costs of survey-based data collection are increasing. The budgetary environment for many statistical agencies is tight. At the same time, traditional surveys may face limitations in capturing emerging features of the economy, such as the growing prevalence of self-employment in the so-called "gig economy." In addition, users of statistical products have increasing demands for information to be more granular—that is, providing more detailed for geographic areas or subgroups of interest—and timelier.

In response to these challenges, statistical agencies are making increased use of alternative data sources to supplement or to substitute for survey data. This report focuses on two key examples of these alternative, non-survey data—administrative data and proprietary data. Other forms of non-survey data, such as image data, sensor data, and certain forms of unstructured "big data" were not components of the products considered here and are beyond the scope of this report.

Government program agencies create administrative data as part of their daily operations. Program agencies at the federal, state and local level collect information on individuals, households, businesses, hospitals, schools, etc. as they administer a government program that involves these entities. A taxonomy by Brackstone (1987) categorizes administrative data based on whether they pertain to: cross-border flows of goods and people; legal registration of certain events (births and deaths, marriage and divorce, business incorporations); administering benefits or obligations (unemployment and health insurance, social assistance programs, taxation); regulation of industry (telecommunications, banking, transportation); and provision of utilities. To support greater evidence to be developed for policymaking, the Commission on Evidence-Based Policymaking developed recommendations for improving researcher access to administrative data while also improving data security and privacy protection (Commission on Evidence-Based Policy, 2017, p. 1).

Private-sector organizations create proprietary data on a wide range of topics. As defined here, *proprietary data* are defined based on ownership: they belong to the private-sector organization that creates them (subject to the rights of those whom the data describe). The use of the data requires a statistical agency to reach an agreement with that organization or meet some type of requirement. Proprietary data may be available to a statistical agency for free or for a fee or other conditions.

As discussed in Chapter 4 on proprietary data, the BLS collects data on prices and expenditures to construct the Consumer Price Index (CPI). Historically, the BLS has focused on collecting such data through surveys and uses the term *alternative data* to refer collectively to non-survey data (Konny et al., 2019, p. 3). BLS has developed a taxonomy of three main types of alternative data. *Corporate data* are "datasets obtained directly from corporate headquarters." *Secondary source data*, also known as *third-party datasets*, are "compiled by a third party, contain prices for goods or services from multiple establishments"; these third-party datasets differ from corporate data because they contain data from more than one organization. *Webscraping data* are "data collected by BLS staff automatically [using software]" or by Application Programming Interfaces (APIs) provided by the firms. Price data collected from a website manually, rather than automatically, are not web scraping data but instead are considered to be "survey" data, like price data collected in-person at a physical outlet.

The strengths and limitations of data vary across types of data sources. One strength of survey data is that a survey can be designed to be nationally representative (or representative of some chosen universe). Another strength is that surveys can collect information on items not available in administrative data (Jarmin, 2019, p. 173). For example, surveys can collect data on socio-economic variables and on program outcomes that can be used for assessing program efficiency and effectiveness. These outcomes and other variables may not be collected in administrative data, which typical include only the information needed to operate the program. A limitation of survey data is that survey responses can contain measurement errors, such as when a person who receives benefits from a government program does not accurately report such receipt to the survey (Meyer et al., 2009). A corresponding strength of a program's administrative data is that, in principle, the data are complete, containing records of the full universe of people who received a program benefit. The administrative data are not typically based on self-reports but on the administration of the program. Administrative data also provide accurate information on duration and amount of benefits each month. It should be noted that administrative data can also be subject to measurement error depending on the data entry method or other errors in accounting. Lastly, a potential limitation of administrative data is that they contain information only on those who participate in the program. A comparison between participants and non-participants would require data besides administrative data.

Administrative data and proprietary data have certain strengths in common. Because these non-survey data are already collected, a statistical agency may be able to acquire them at a cost that is lower than the cost of fielding a new survey to collect equivalent data. In addition, re-using non-survey data for statistical purposes can reduce burden on the American public compared to fielding a new survey (Office of Management and Budget, 2014, p. 7). Depending on what data sources are considered, non-survey data can be more detailed than survey data with respect to program receipt and benefit utilization.

Countering those strengths of non-survey data, however, are other issues of data quality. A major example of such an issue is coverage—the extent to which the data are drawn from, and thereby represent, a target population or universe of interest to the statistical agency (which can have different populations of interest than those of the program agency or business that created the non-survey data). Non-survey data may be available for only a portion of the universe of interest. For example, administrative data cover only program participants, excluding non-participants for whom information may be needed for comparative statistical analysis. Similarly, proprietary data for, say, retail prices of various foods may cover only some portion of retailers—prices from large supermarket chains may be included while prices from some smaller, independent groceries are not. (Muth, 2016 p. 41) In contrast, a survey's sample frame can, in principle, contain program participants and non-participants alike. It can include all types of retailers, or hospitals, or schools, etc. Because survey data are designed specifically for statistical purposes, survey data can be designed to cover the universe of interest to support estimation of nationally representative statistics in a way that non-survey data can find challenging.

When a statistical agency combines data from multiple sources into a single dataset, the result is integrated data, which may also be known as blended, hybrid or mixed data. Precisely because there are strengths and limitations from using any single source of data—whether it is survey or non-survey data—there is a unique advantage when statistical analysis integrates multiple sources of data. As explained in OMB guidance in the context of administrative data, "The ability to combine administrative datasets with each other or with survey data offers significant potential to answer important questions that neither type of data can answer alone—questions whose answers may be particularly applicable to program agencies seeking to increase program efficiency and efficacy" (Office of Management and Budget, 2014, p. 7). For example, variables in one data source may not be available in another. The strengths of the different sources can be leveraged by combining them; equivalently, it can be said that different data sources can have different limitations or types of errors that can potentially compensate for one another by combining them. At the same time, there is no guarantee that quality always improves when using multiple sources of data—errors can compound one another rather than compensate.

On a case-by-case basis, statistical agencies make decisions on whether and how to integrate data. When disseminating a statistical product based on integrated data, agencies inform users about the sources of data, how the sources were integrated, and the possible implications for data quality. The content and details of such reporting can influence a users' understanding of data quality and how to best use and interpret the results for their own applications.

Dimensions of Data Quality. High-quality information and appropriate reporting by federal agencies, including statistical agencies, is a priority for the U.S. Congress, the Office of Management and Budget and the agencies themselves. In response to the Treasury and General Government Appropriations Act for Fiscal Year 2001, OMB issued guidelines that, in part, direct each Federal agency to issue its own implementing guidelines ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by the agency (67 Federal 8452-8460). The OMB guidelines are sufficiently general to apply across the highly diverse types of information disseminated by federal agencies. The need for agencies to tailor the broad OMB guidelines to their own circumstances was recognized by the OMB guidelines, which state that, "while agencies' implementation of the guidelines may differ, the essence of the guidelines will apply" (67 Federal Register, p. 8453). Each agency considers how best to meet the OMB guidelines to ensure and maximize the quality, objectivity, utility, and integrity of information. A discussion of OMB guidelines that are related to transparent reporting is in National Academies of Sciences, Engineering, and Medicine (2019, pp. 8-12).

The discussion of *utility* in the OMB guideline embodies a user's perspective: "In assessing the usefulness of information that the agency disseminates to the public, the agency needs to consider the uses of the information not only from the perspective of the agency but also from the *perspective of the public*" (67 Federal Register, p. 8459, emphasis added). Similarly, the most general and widely quoted definition of data quality in the statistical literature may be "fitness-for-use" (Biemer and Lyberg, 2003, p. 13). Thus, if some feature of the data matters to users, affecting the fitness-for-use, then that same feature is tantamount to a feature of data quality.

Transparent reporting on a broad range of features of data quality enables users to make an informed, user-specific assessment of quality. The data quality framework used in the Transparent Reporting Project was based largely on a framework adopted internally at BEA, one of the seven agencies participating in the project. This framework consists of eight dimensions: *relevance*, *accuracy*, *reliability*, *timeliness*, *punctuality*, *consistency*, *comparability*, and *access*. This set of eight dimensions reflects the influence of various data quality frameworks used in the United States and elsewhere, including those laid out by the Organization for Economic Cooperation and Development (Organization for Economic Cooperation and Development, 2015, p. 20), which in turn drew from the quality assurance framework of the European Statistical System (ESS).

The comprehensive FCSM report *Measuring and Reporting Sources of Error in Surveys* (Federal Committee on Statistical Methodology, 2001) discussed in detail the meaning of several dimensions of data quality, and provided brief definitions of other dimensions drawn from the literature, as shown in table 1.2.

[Table follows on next page]

Table 1.2. Dimensions of Data Quality

Relevance refers to the idea that the data collection program measures concepts that are meaningful and useful to data users. Does the concept implemented in the data collection program fit the intended use? For example, concepts first measured in a continuous sample survey program 20 years ago may be inapplicable in current society; that is, it may no longer be relevant to data users. Determining the relevance of concepts and definitions is a difficult and time-consuming process requiring the expertise of data collectors, data providers, data users, agency researchers, and expert panels.
Accuracy is an important and visible aspect of quality that has been of concern to statisticians and survey methodologists for many years. It relates to the closeness between estimated and true (unknown) values. For many, accuracy means the measurement and reporting of estimates of sampling error for sample survey programs, but, in fact, the concept is much broader, taking in nonsampling error as well. Nonsampling error includes coverage error, measurement error, nonresponse error, and processing error … [I]t is important to recognize that the accuracy of any estimate is affected by both sampling and nonsampling error.
Timeliness can refer to several concepts. First, it refers to the length of the data collection's production time—the time from data collection until the first availability of a product. Fast release times are without exception looked upon favorably by end users. Second, timeliness can also refer to the frequency of the data collection. Timely data are current data. Timeliness can be difficult to characterize since the characteristics of the data collection can affect the availability of data. For example, a new sample survey may require more time prior to implementation than the revision of an existing survey. Data from continuous recurring surveys should be available sooner than periodic or one-time surveys, but ultimately timeliness is assessed by user needs and expectations.
Accessibility, as a characteristic of data quality, refers to the ability of data users to obtain the products of the data collection program. Data products have their most value—are most accessible—when they are easily available to end-users and in the forms and formats desired. Data products are of several types—individual microdata in user-friendly formats on different media, statistical tabulations on key survey variables, and analytic and descriptive analysis reports. Accessibility also implies the data products include adequate documentation and discussion to allow proper interpretation of the survey results. Accessibility can also be described in terms of the efforts data producers make to provide "hands-on" technical assistance in using and interpreting the data products through consultation, training classes, etc.
Comparability of statistics refers to the ability to make reliable comparisons over time.
Coherence refers to the ability of the statistical data program to maintain common definitions, classifications, and methodological standards when data originate from several sources.
Completeness is the ability of the statistical data collection to provide statistics for all domains identified by the user community.

Source: Federal Committee on Statistical Methodology (2001, pp. 1-2 – 1-3)

These quality dimensions are interrelated and may to some extent overlap one another. Some dimensions may be quantitative and lend themselves to metrics and measurement. Others are qualitative—they can be described and discussed but metrics are lacking. While FCSM (2001) introduced various dimensions of quality, as conveyed in the table above, the report focused on accuracy and five sources of error that affect accuracy (Federal Committee of Statistical Methodology, 2001, pp. 1-3, 1-5).

While these dimensions originated in the context of survey data quality, they have been extended to apply to administrative data and integrated data. For example, a previous ad hoc subcommittee of the FCSM adopted many of these dimensions when it developed a Data Quality Assessment Tool for statistical agencies to use as they consider acquisition of administrative data (Iwig et al., 2013). These dimensions have been discussed in the report *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* written by an expert panel convened by the Committee on National Statistics (CNSTAT) of the National Academies of Science, Engineering, and Medicine (National Academies of Sciences, Engineering, and Medicine, 2017b, pp. 114-115). For these dimensions, Appendix 1 provides the CNSTAT report's definitions and complementary definitions from by the ESS's *Handbook on Data Quality Assessment Methods and Tools* (Ehling and Körner, 2007). Definitions in these sources are proximate with those given in FCSM (2001).

In two ways, this report considers issues of reporting on data quality that go beyond the eight core dimensions. First, the report's third chapter reviews how data quality is assessed at the Department of Veterans Affairs (VA), of which the National Center for Veterans Analysis and Statistics (NCVAS) is a part. It is helpful to convey how an agency that concentrates largely on administrative data addresses issues of data quality.

Second, the report goes beyond the eight core dimensions by giving some attention to granularity— the level of detail available for statistics on geographic areas or subgroups. The CNSTAT report emphasized that *granularity* and *timeliness* are two dimensions that may be undervalued (National Academies of Sciences, Engineering, and Medicine, 2017b, pp., 116-117). A strength of using non-survey data sources as part of an integrated data strategy is the potential improvement in *granularity*, an aspect of data quality that can be given more prominence by making it an additional dimension in a data quality framework.

On the other hand, *granularity* can be considered to be an integral aspect of *relevance*. Another expert panel, convened by CNSTAT to examine the Census Bureau's annual economic surveys, stated in its report that "Relevance may reflect a variety of dimensions, including the availability of data that are *sufficiently detailed* to monitor changes in the economy and meet policy needs, and the availability of data to address the full range of topics important for public and private understanding and decision making." (National Academies of Sciences, Engineering, and Medicine, 2005, p. 22, emphasis added) Whether *granularity* is treated as a stand-alone dimension or subsumed within *relevance* or another dimension, transparent reporting calls for *granularity* and other salient aspects of quality to be addressed in agency documentation. The ICSP has developed a set of Principles for Modernizing Production of Federal Statistics, which state, "While fully complying with confidentiality and privacy requirements, agencies should continue to make statistical information created in support of mission activities as granular and timely as practicable and widely accessible" (Interagency Council on Statistical Policy, 2018, p. 3).

The eight dimensions in the quality framework used by the Transparent Reporting Project cover many key components of quality. The project incorporated these dimensions into its customer survey, which was developed in mid-2018. The ICSP principles, which were issued after the customer survey was developed, identify additional components of transparent public reporting that were not explicitly included in the survey. For example, the ICSP principles include integrity (as do the OMB guidelines), but the customer survey did not ask users specifically about it. The examination of quality conducted by the Transparent Reporting Project is detailed but not exhaustive.

Methodology of the Customer Survey. The customer survey was designed to enable users who participated in the survey to distinguish between data quality—whether the quality of the data is high or low (by a given dimension)—and the thoroughness of the explanation provided by the statistical agency about that quality. An agency provides transparent reporting by describing the product's data quality with detail that is enough for the user to make an informed assessment about fitness for use. Agency reporting can be transparent by discussing threats to data quality, which can impede fitness for use. Such threats may be discussed in terms of "sources of errors" or "limitations of the data." The ICSP principles state "[S]imilar to reporting on statistical information produced from traditional sources, if there are known deficiencies or limitations in products produced from a non-statistical or integrated data source, these should be clearly articulated." (Interagency Council on Statistical Policy, 2018, p. 3).

Over the years, the agencies in the Transparent Reporting Project have each been in touch with stakeholders of their statistical products. The project is far from the first occasion at which the agencies learn about their stakeholders' assessment of the product and its documentation. Even so, the project provided new opportunities for the agencies. First, the agencies could learn about user assessments that focused heavily on documentation. Second, the agencies could obtain a more current snapshot of user assessments (as of late 2018) than provided by more dated feedback. Third, perhaps most importantly, agencies could obtain user assessments using a common, cross-agency customer survey.

The cross-agency nature of the project was an important feature of the customer survey: it is the first known survey to collect user assessments of transparent reporting across integrated-data statistical products produced by multiple agencies, all of which were assessed using common items in a common framework. For such a survey, the project developed questions that transcended products, topics or documents. As a result, the project's cross-agency approach provides a broader perspective on the agency's products and documentation than what might be achievable in an environment that focused on a single product only. That feature of the project is useful for the participating agencies. In addition, by moving beyond items that have terms of reference tied to one agency, the survey can serve as an example for many agencies who can draw upon the survey for their own inquires with their stakeholders.

This effort builds on previous cross-agency collaboration and research and examining reporting practices. For example, the FCSM report Measuring and Reporting the Quality of Survey Data conducted studies that examined how various publications of 12 statistical agencies reported sources of error in survey data (Federal Committee on Statistical Methodology, 2001, p. 2-2). The work of the Transparent Reporting Project differs from the previous FCSM work in scope, because we consider integrated statistical products rather than focus on survey-based products. The project's work also differs in methodology due to the inclusion of the customer survey.

The project faced three challenges in developing the customer survey. First, a cross-agency environment is a challenge because items must be written in a sufficiently general way that they can be answered by users of a variety of products. Cross-agency work is inherently complex. Ultimately, depending on level of specificity, not all of a survey's items may be applicable for each integrated data product. For example, items about linkage procedures may not be salient for national accounts.

Second, the project had to develop items from first principles. There was no pre-existing standard instrument for asking detailed items about users' assessments of agency documentation of statistical products. The survey covered several topics, including: general assessment; the product's data sources; evaluation of the quality of the product; how to use the product; users contacting agency staff for information about the product; and suggestions for improving data quality for the product.

Third, users of statistical products have diverse needs. This key point has been recognized previously. For example, a CNSTAT review of SESTAT conducted for NCSES stated "users of the data tables may be very different from users who typically download the microdata." (National Academies of Science, Engineering, and Medicine, 2018, p. 40) At one of the workshops hosted in 2018 by the FCSM Working Group and the Washington Statistical Society, three categories of transparency were identified that correspond to different types of users: "High Transparency for academics, agency specialists, subject-matter experts; Moderate Transparency for policy makers, professional journalists, students; Low Transparency for the general public" (Federal Committee on Statistical Methodology Working Group, 2018, p. 9). Influenced by the FCSM Working Group's identification that a user typology can be helpful, the Transparent Reporting Project built that approach into its customer survey.

A second feature of the project's customer survey is that it contained two instruments, tailoring each instrument for the two types of users who were contacted to participate in the survey. The agencies in the project gave the name “informed consumers” to one of the groups of users. Informed consumers use results based on the integrated data product and need information on how the product was developed to interpret the information it provides. The other group was named “researchers”—those who either directly use raw data in the integrated data product or need detailed technical information about the product for their research activities. The project’s two instruments shared many items in common. Other items appeared on only one of the two instruments. There was opportunity for agencies in the project to add a few, highly targeted agency-specific items that would be administered only to users of their product.

The informed consumer instrument contained 56 items, of which 2 were agency-specific items that were administered to a subset of informed consumers; see Appendix 2. The researcher instrument contained 62 items, of which 3 were agency-specific items (2 of which were also on the informed consumer instrument); see Appendix 3. Apart from the 3 agency-specific items, the two instruments had 48 items in common. Altogether, there were 68 distinct items: 48 common items, 6 additional items that appeared only on the informed consumer instrument, 11 items that appeared only on the researcher instrument, and 3 agency-specific items (that appeared on one or both instruments). For any one user, there were either 56 or 62 items.

To garner some kinds of information, statistical agencies solicit detailed feedback from key stakeholders who are known to the agency. The Transparency Reporting Project followed that practice in its administration of the customer survey. Specifically, the project used purposive (or purposeful) sampling rather than a probability sample; a probability survey would have been infeasible, even if desired, because the project lacked a sampling frame of the full population of all users of a statistical product. Instead, the people contacted to participate in the survey were chosen by the agencies as groups of users who have a familiarity with the product and a stake in the product and its documentation. Such users may be especially able to provide detailed responses and they may be more willing to participate in a survey than a user selected completely at random. In general, purposive sampling “is a technique widely used in qualitative research for the identification and selection of information-rich cases” which involves “identifying and selecting individuals or groups of individuals that are especially knowledgeable about or experienced with a phenomenon of interest” (Palinkas et al., 2015, p. 2). That is, a purposive sample is formed when an organization itself selects known stakeholders, rather than using a probability sample, to provide qualitative or quantitative information.

The substantial limitation of purposive sampling is that the sample is not representative in a statistical sense. Therefore, results from the project’s customer survey cannot be generalized to make inferences about the population of all users. Nevertheless, the results can be valuable and informative about how certain (selected) users assess how well an agency reports on its data quality. These results can be suggestive of what the broader population of users may be thinking. The survey and its responses represent the most current information available on user assessments of agency products and documentation and the best information available on a cross-agency basis.

Agencies on the project selected possible survey participants based their knowledge of customers who may be especially familiar with the statistical product or who use it in their work. The agencies developed two lists of selected users, distinguished by type of customer, who were contacted to participate in the survey through their work e-mails. The difference between informed consumers and researchers reflects the extent to which the user interacts with the statistical product—not the user's place of work or occupational title. There was overlap in the types of organizations that employ the two types of customers. Informed consumers contacted for the survey included university academics, experts working at private companies, and government specialists at program agencies that make use of the statistical product. Contacted researchers included university academics, staff at the Congressional Budget Office and the Congressional Research Service, government staff at program agencies, and researchers at non-profit organizations. Some of the users contacted for the customer survey could arguably be in either of the two groups.

Survey participants each rated only one of the statistical products and its documentation. Because the BLS reports on proprietary data on telecommunications prices were available internally to BLS staff and not the general public, there were no external users to contact to assess the BLS documentation; despite that difference with other agencies' products, the BLS product was included in the Transparent Reporting Project because it was anticipated that a review of BLS documentation by the project would strengthen the project and be helpful for readers of the project report.

The survey was administered through Grant Thornton (GT), a private consulting firm that was engaged in the project through the General Services Administration (GSA). To develop and administer the survey, the agencies on the project worked in tandem with staff from GT and GSA, who were helpful in rewording certain items and in developing web-based instruments. A web-based data collection enabled the survey to record user responses automatically. User responses were made by either radio buttons or typewritten text, depending on the item. NCES represented the Transparent Reporting Project in the agency coordination of OMB review and clearance of the proposed items. Via e-mail, agencies themselves notified users in advance that the users would receive a survey from GSA and encouraged users to participate. Users later received e-mails with a link to the survey and e-mail reminders to complete the survey as needed. The survey was open to users from November 1 through December 10, 2018.

Response rates for the customer survey. This section examines responses rates for the Transparent Reporting Project's customer survey. Not all products were reviewed by both informed consumers and researchers, largely reflecting agency discretion on whether to invite one or both types of customers to participate. For example, some statistical products tend to be used most intensively by researchers, making informed consumers difficult to identify.

As shown in Table 1.3, the project sent in total 105 surveys to users who were asked to participate in the survey. Of these, 46 participated in the survey for a response rate of about 44 percent (46/105).

Table 1.3. Responses to Transparent Reporting Project Customer Survey, by Type of Customer and by Product

Type of Customer	Product	Agency	Number of surveys completed	Number of surveys sent	Response rate (percent)
Informed Consumer					
	PCE	BEA	1	6	16.7
	IRI Data	ERS	0	2	0.0
	SESTAT	NCSES	1	6	16.7
	VetPop2016	NCVAS	6	17	35.3
		Subtotal	**8**	**31**	**25.8**
Researcher					
	FoodAPS	ERS	5	18	27.8
	IRI Data	ERS	5	7	71.4
	NPSAS	NCES	14	21	66.7
	NCHS	NCHS	7	8	87.5
	SESTAT	NCSES	3	11	27.3
	VetPop2016	NCVAS	4	9	44.4
		Subtotal	**38**	**74**	**51.4**
Total			**46**	**105**	**43.8**

Source: Transparent Reporting Project Customer Survey

The customer survey's overall response rate is lower than on some major federal surveys. For example, in October 2018 the Current Population Survey had a response rate of 84.8 percent (Bureau of Labor Statistics, 2019a). In 2013-15, seven health surveys conducted by the U.S. Department of Health and Human Services (HHS) each had response rates above 70 percent (Czajka and Beyler, 2016). However, in fiscal 2012, when hospitals conducted the customer satisfaction survey known as the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS, a standardized 27-item survey designed by two agencies in HHS), the response rates averaged 32 percent—less than half the response rate for the seven HHS health surveys (Siegrist, 2013, p. 983). Moreover, that response rate of 32 percent for the HCAHPS was “substantially higher” compared to response rates to customer satisfaction surveys in most industries, perhaps due to the “relative importance of the health care experience.” Thus, the 44-percent response rate of the Transparency Reporting Project’s customer survey exceeds the typical response rates for most customer surveys.

There was a notable difference in response rate between the two subgroups of users. Only 8 of the 31 informed consumers contacted for the survey chose to participate, representing a response rate of about 25.8 percent. Researchers participated at a rate of 51.3 percent, with 38 researchers participating out of 74 contacted. The difference between the subgroups’ response rates was just over 25 percentage points; alternatively, the response rate of the researchers was about double the rate of the informed consumers.

It may be speculated that researchers have relatively higher response rates because they have an especially close connection with the statistical product. However, informed consumers can rely heavily on the estimates or tables an agency produces. For at least some informed consumers, there can be a strong interest in helping an agency improve its documentation. In the end, it is unknown why the response rates for the two groups differed.

The relatively low response rate of informed consumers affected the pattern of responses across the agencies differently. For example, BEA used the project’s customer survey as an opportunity to reach out to informed consumers, soliciting participation from six such users of whom only one chose to participate; prior to the survey, it was not known what the response rates of informed consumers or researchers would be. Given the low response rate of informed consumers and that BEA focused on that subgroup, the agency sample size for BEA becomes understandable. Like BEA, NCSES contacted six informed consumers and had only one who participated. The ERS experience resembledthat of the other two agencies, with ERS contacting two informed consumers to participate and neither doing so. NCVAS had an exceptionally good experience with informed consumers: of the 17 who were contacted, 6 chose to participate.

It is important to bear in mind that the Transparent Reporting Project’s customer survey was meant to be exploratory and illustrative—a first attempt at soliciting user responses on this complex subject. It was determined that it was important to get an initial set of user responses so that the responses could be evaluated and the reporting process improved. A more refined data collection could have taken substantially more time.

Data Quality and Confidence in the Data. The focus of the survey was on assessing agency documentation. Even so, the customer survey was an opportunity for the agencies to learn about user appraisals of the data quality of their statistical products and of users “overall confidence” in the data. The customer survey opened with these items so that users could then turn their attention to discussing agency documentation about the data. Not all survey respondents provided a response to every item that was available to them.

The 8 questions on data quality and confidence in the data are provided below in table 1.4; appendix 2 for informed consumers and appendix 3 for researchers provide the full wording for all items on the customer survey.

Table 1.4. Items on Data Quality and Confidence in the Data on the Customer Survey

These next questions refer to various dimensions of quality of {Product}. Please rate your satisfaction using a 5-point scale to respond where "1" is "Very dissatisfied," "2" is "Somewhat dissatisfied," "3" is "Neither dissatisfied nor satisfied," "4" is "Somewhat satisfied," and "5" is "Very satisfied." If you do not know the answer or an item does not apply to you, please select "N/A" (not applicable).

How would you rate your satisfaction with {Product} in terms of:

A3. Relevance to your research or reporting needs

A4. Accuracy of information. That is, does it effectively measure the issue for which you need data

A5. Whether the information is reliable in terms of being based on scientific criteria used to selected data sources and statistical methods

A6. Time between when information in {Product} was collected and when it was available to you

A7. Time between when the information was scheduled to be available to you (target date) and the time it actually became available

A8. Consistency with other information that you know about the topic you needed {Product} to study

A9. Comparability, or whether information about a topic from one source of information in {Product} was comparable to information about that topic from another source of data used in {Product} (e.g., pricing data from different sources were for comparable units of a product). This includes comparability to previous releases of {Product}.

A10. Ease of accessing {Product}

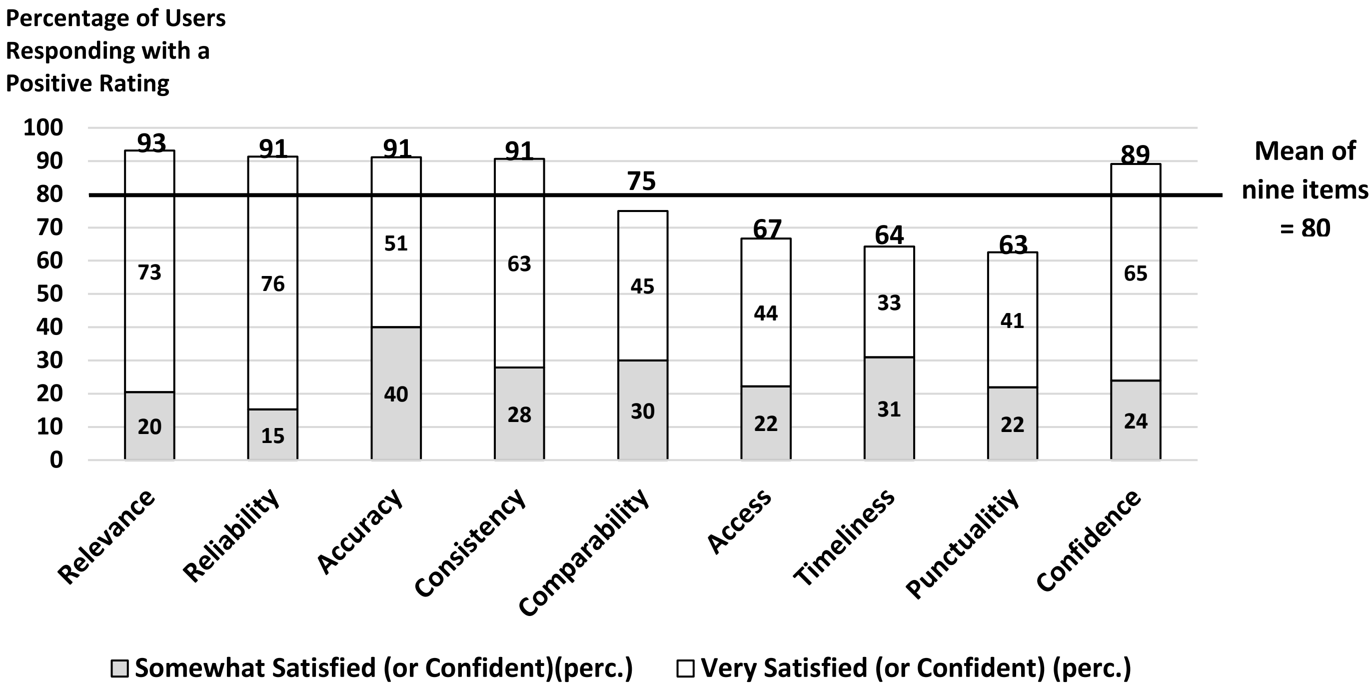
A11. Overall, how confident are you in the data or the statistics that you obtain from {Product}? (1=Very unconfident/2=Somewhat unconfident/3=Neither unconfident nor confident/4=Somewhat confident/5=Very confident)

Source: Transparent Reporting Project Customer Survey

Appendix 4 shows, by item, the number of users who responded (col. 5), the number of users with a positive rating of either 4 or 5 (col. 7), and the percentage of users with a positive rating (col. 8); the appendix also reports numerical means by user type, for informed consumers and researchers (columns 11 and 12). Appendix 5 reports numerical means further disaggregated by user type and by statistical product. In this section we consider the aggregate results.

Among the 44 users who responded to the item about relevance, 41 of them answered with either a 4 (Somewhat satisfied) or a 5 (Very Satisfied) on the 5-point Likert scale, that is, 93 percent of users (who responded to the item) had a positive rating, reflecting that 20 percent had a rating of Somewhat Satisfied (or Confident) and 73 had a rating of Very Satisfied (or Confident). The percentages of users who had a positive rating for the various dimensions of data quality are shown in Figure 1.1.

Figure. 1.1. User Assessments for Dimensions of Data Quality and User Overall Confidence in the Data



Note: A positive rating is a response of "Somewhat satisfied" or "Very Satisfied" on a Likert scale assessing each dimension of data quality and "Somewhat confident" or "Very confident" when describing overall confidence in the data. A percentage of positive ratings is displayed above each box.
Source: Transparent Reporting Project Customer Survey.

Besides the result for relevance, the figure shows that the dimensions of accuracy reliability, and consistency also had relatively high percentages of positive ratings, with 91 percent of users responding with a 4 or a 5.

In contrast, another group of dimensions had positive ratings that were relatively lower. The data quality dimensions with the lowest percentages of positive ratings were timeliness and punctuality, with 64 and 63 percent of users, respectively, responding with a 4 or 5; it is noted that only 32 users responded to the punctuality item, which may indicate users that some did not consider this dimension to be applicable to their statistical product.

Access was slightly better than timeliness and punctuality, with 67 percent of users rating it positively. Higher still was comparability, at 75 percent. Nevertheless, this group of four dimension have their highest percentage of positive ratings at 75 percent, while the percentages exhibited by relevance, accuracy, reliability, and consistency are all over 90 percent.

A bottom-line assessment of data quality is captured by a user assessment of confidence in the data or statistics. The percentage of positive ratings was relatively high at 89 percent.

The mean for the item on overall assessment of confidence and the eight items on data quality is about 80 percent; as an unweighted mean, that figure does not consider that different numbers of users responded to the nine items.

While the origins of the Likert scale make it an ordinal measure, sometimes Likert responses are interpreted numerically. Proceeding with that interpretation allows for a mean response to be numerically calculated for an item. Historically, different agencies that solicit user assessments with some type of Likert scale have used either a percentage of users or a mean as an indicator of user satisfaction. For example, NCES has examined the percent of surveyed customers that "agree that the NCES data (publications and data files) are timely, relevant, and comprehensive" (Lemke et al., 2001, p. 4). In contrast, the BEA has used a "customer satisfaction rating" as an indicator, which use an item's mean response. Thus, the percentage of positive ratings and the numerical mean have each been used as indicators.

Proceeding with a numerical interpretation of Likert responses, the distribution of numerical responses for relevance (one 2; two 3's; nine 4's; and thirty-two 5's) results in a numerical average of about 4.6. The numerical means of Likert items and of dichotomous items are provided in Appendix 4 (col. 9), along with the standard deviation of an item's responses (col. 9). As was true with the percentages of users with high ratings, the values of the overall means for the quality dimensions fall into two clusters. Relevance, accuracy, reliability and consistency have relatively high means, in the range of 4.4 to 4.6 on a 5-point scale. The numerical mean for confidence is also relatively high at 4.5. In contrast, timeliness, punctuality, comparability and access have relatively low means, ranging from 3.8 to 3.9.

For these nine items, the numerical values of the two indicators—the percentage of positive ratings and the mean—have a high correlation coefficient of 0.98. Thus, in this case, the information content of the two measures are essentially the same: what either measure tells about user assessments is mirrored by the other.

Appendix 4 reports numerical means by user type, for informed consumers and researchers (columns 11 and 12). Appendix 5 reports numerical means further disaggregated by user type and by statistical product. At that level of granularity, agency sample sizes are small, and the results may not be robust. Of course, user responses are most numerous when considering all the respondents together as a single group. That approach is adopted in the body of the report, with discussion of differences by type of user reserved for Appendix 6.

Finally, to reiterate, we recognize that the small non-probability sample and number of responses limit the strength of any inferences. However, we believe that the results can serve as useful indicators for agencies to examine their current practices.

Remainder of the report. The remainder of the report turns from issues of data quality per se to the core of the report—issues of transparent reporting about data quality. The next four chapters each consider a type of integrated data in detail. These chapters introduce the case studies for several statistical products, explain the products' use of integrated data, analyze how the documentation communicates data quality to users, provide a rough estimate of the associated cost for the agencies, and highlight user product-specific assessments. Chapter 6 analyzes user assessments of documentation for all the products under study as a group. Chapter 7 draws upon findings in previous chapters to identify challenges to transparent reporting, and Chapter 8 provides general conclusions.

References for Chapter 1 (Introduction)

Biemer and Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons.

Brackstone, G.J. 1987. "Issues in the Use of Administrative Records for Statistical Purposes," *Survey Methodology* Vol. 13, No. 1, pp. 29-43.
https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198700114467

Commission on Evidence-Based Policymaking. 2017. *The Promise of Evidence-Based Policymaking.* September, 2017.
https://www.cep.gov/

Czajka, John and Amy Beyler. 2016. *Declining Response Rates in Federal Surveys: Trends and Implications*, prepared by Mathematica Policy Research for U.S. Department of Health and Human Services, Office of the Secretary for Planning and Evaluation. Washington DC: Mathematica Policy Research. June 2016.
https://aspe.hhs.gov/system/files/pdf/255531/Decliningresponserates.pdf

Economic Research Service. 2019. "ERS Data Product Quality Standards," webpage updated April 1, 2019
https://www.ers.usda.gov/about-ers/policies-and-standards/data-product-quality/ers-data-product-quality-standards/

Ehling, Manfred, and Thomas Körner, eds. 2007. *Handbook on Data Quality Assessment Methods and Tools*. Eurostat, European Commission. 2007.
https://ec.europa.eu/eurostat/web/quality/quality-reporting

Federal Committee on Statistical Methodology Subcommittee on Measuring and Reporting Sources of Errors in Surveys. 2001. *Measuring and Reporting Sources of Error in Surveys.* Working Paper 31.
https://nces.ed.gov/FCSM/pdf/spwp31.pdf

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

Federal Register, Vol. 67, No. 36, February 22, 2002, pp. 8452-8460.
https://www.gpo.gov/fdsys/pkg/FR-2002-02-22/pdf/R2-59.pdf

Interagency Council on Statistical Policy. 2018. *Principles for Modernizing Production of Federal Statistics.*
https://nces.ed.gov/fcsm/pdf/Principles.pdf

Iwig, William, Michael Berning, Paul Marck, and Mark Prell. 2013. *Data Quality Assessment Tool for Administrative Data*. Federal Committee on Statistical Methodology Working Paper No. 46.
https://nces.ed.gov/FCSM/policy_wp.asp

Jarmin, Ron. 2019. "Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics." *Journal of Economic Perspectives.* Vol. 33, Nol. 1. Winter 2019.
https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.33.1.165

Konny, Crystal, Brendan Williams, and David Friedman. 2019. "Big Data in the U.S. Consumer Prices Index: Experiences & Plans," paper for Brookings Institution panel "Can big data improve economic measurement?"
https://www.brookings.edu/wp-content/uploads/2019/02/Big-Data-in-the-U.S.-Consumer-Price-Index.pdf

Lemke, Mariann, Val Plisko, Marilyn Seastrom, and Daniel Kasprzyk. 2001. "Benchmarking the Performance of Statistical Agencies," in *Proceedings of Statistics Canada Symposium 2001—Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. available at https://www150.statcan.gc.ca/n1/pub/11-522-x/11-522-x2001001-eng.htm

Meyer, Bruce, Wallace Mok and James Sullivan. 2009. *The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences*. National Bureau of Economic Research Working Paper 15181, July 2009. https://www.nber.org/papers/w15181

Muth, Mary, Megan Sweitzer, Derick Brown, Kristen Capogrossi, Shawn Karns, David Levin, Abigail Okrent, Peter Siegel, and Chen Zhen. 2016. *Understanding IRI Household-Based and Store-Based Scanner Data.* Technical Bulletin No. 1942. USDA, Economic Research Service. https://www.ers.usda.gov/publications/pub-details/?pubid=47636

National Academies of Sciences, Engineering, and Medicine. 2005. *Reengineering the Census Bureau's Annual Economic Surveys.* Washington, DC: The National Academies Press. https://www.nap.edu/catalog/25098/reengineering-the-census-bureaus-annual-economic-surveys

National Academies of Sciences, Engineering, and Medicine. 2017a. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/24652/innovations-in-federal-statistics-combining-data-sources-while-protecting-privacy

National Academies of Sciences, Engineering, and Medicine. 2017b. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/24893/federal-statistics-multiple-data-sources-and-privacy-protection-next-steps

National Academies of Sciences, Engineering, and Medicine. 2018. *Measuring the 21st Century Science and Engineering Workforce Population: Evolving Needs*. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/24968/measuring-the-21st-century-science-and-engineering-workforce-population-evolving

National Academies of Sciences, Engineering, and Medicine. 2019. *Methods to Foster Transparency and Reproducibility of Federal Statistics: Proceedings of a Workshop*. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/25305/methods-to-foster-transparency-and-reproducibility-of-federal-statistics-proceedings

Organization for Economic Cooperation and Development. 2015. *Recommendations of the OECD Council on Good Statistical Practice*. http://www.oecd.org/statistics/good-practice-toolkit/Brochure-Good-Stat-Practices.pdf

Palinkas, Lawrence, Sarah Horwitz, Carla Green, Jennifer Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," *Administration and Policy in Mental Health and Mental Health Services Research*. September 2015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012002/

Siegrist, Richard. 2013. "Patient Satisfaction: History, Myths, and Misperceptions," *American Medical Association Journal of Ethics*. Vol. 15, No. 11. November 2013. https://journalofethics.ama-assn.org/article/patient-satisfaction-history-myths-and-misperceptions/2013-11

U.S. Office of Management and Budget. 2018. Statistical Programs of the United States
https://www.whitehouse.gov/wp-content/uploads/2018/05/statistical-programs-2018.pdf

U.S. Office of Management and Budget. 2014. Guidance for Providing and Using Administrative Data for Statistical Purposes, Memorandum M-14-06.
https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf

U.S. Office of Management and Budget. 2019. Improving Implementation of the Information Quality Act, Memorandum M-19-015.
https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf

2. National Accounts

— Personal Consumption Expenditures component of Gross Domestic Product —

Bureau of Economic Analysis

The National Income and Product Accounts (NIPAs) produced by the Bureau of Economic Analysis (BEA) provide a picture of the U.S. economy that is timely, comprehensive and accurate (Fixler et al., 2018, p. 1). A key measure of economic activity featured in the NIPAs is Gross Domestic Product (GDP)—the dollar value of expenditures on final goods and services. The NIPAs and GDP estimates are used by government policymakers, business decision-makers, academic researchers, and the public for many purposes (Bureau of Economic Analysis, 2017a, pp. 1-6 – 1-7). They provide indicators of current economic conditions and inputs into forecasts of future conditions. They are used by the White House and Congress in developing federal budgets and fiscal policy and by the Federal Reserve Board in developing monetary policy. They support examination of relationships between sectors of the U.S. economy and comparisons between the U.S. economy and other economies. In 2000, a BEA review of GDP measurement in the 20th century noted that "The national accounts, in combination with better informed policies and institutions, have contributed to a reduction in the severity of business cycles and a post-war era [following World War II] of strong economic growth (Bureau of Economic Analysis, 2000, p. 8). The review quoted Paul Samuelson and William Nordhaus, two winners of the Nobel prize in economics, who wrote, "While the GDP and the rest of the national income accounts may seem to be arcane concepts, they are truly among the great inventions of the twentieth century" (Bureau of Economic Analysis, 2000, p. 6).

Over the span of decades, BEA and extramural researchers have contributed to a large literature on the NIPAs and GDP, each of which are complex, large-scale statistical products; there are about 300 tables that support the NIPAs (Bureau of Economic Analysis, 2017a, p. 1-2) The Transparent Reporting Project could not undertake an assessment of how well BEA documentation describes the NIPAs or GDP as a key statistic, nor ask users in a customer survey to assess them because of the scope and complexity of the GDP measure. To manage the report's scope, the project team choose to focus attention where possible on measurement of the Personal Consumption Expenditures (PCE) component in GDP.

Total expenditures in GDP are categorized into four broad components: PCE, gross private domestic investment, government purchases, and net exports. The expenditure measure of GDP equals the sum of the four components. The project team choose PCE as a focus because PCE is the largest component of GDP, representing about two-thirds (69.4 percent) of the U.S. economy according to preliminary figures for 2018 (Council of Economic Advisers, 2019, p. 634). Questions that can be addressed using PCE estimates include: "How strong was consumer spending compared with the month before? What types of goods or services saw a rise in spending? How did a sharp increase in food or energy prices affect consumers spending?" (Bureau of Economic Analysis, 2018).

The Statistical Product and its Data Sources. The NIPAs and GDP represent integrated data products that combine multiple sources. The PCE component of GDP, conceived of as a stand-alone product for the Transparent Reporting Project, is an integrated data product as well. To estimate PCE, BEA receives data from many government and non-government sources that are listed in Table 2.1.

Table 2.1. Main Source data used in PCE (annual updates)

Item	Data Source Category	Data Source
Goods and Food Services (except as separately specified below)	Census	Monthly Retail Trade Survey (MRTS) Annual Retail Trade Survey (ARTS) Economic Census
Exceptions:		
New motor vehicles	Private/trade	Ward's Automotive Reports (unit sales) J.D. Power and Associates (registrations) R.L. Polk (average prices)
Food produced and consumed on farms	Government or admin	U.S. Department of Agriculture
Standard clothing issued military	Government or admin	Federal budget data
Nondiesel gasoline	Government or admin; Census	Energy Information Administration (quantities) Bureau of Labor Statistics (average prices) Economic Census
Prescription drugs	Private/trade; Census	Intercontinental Medical Statistics Health, Inc. (sales) Economic Census
Tobacco	Government or admin; Census	U.S. Department of Treasury (consumption) Bureau of Labor Statistics (Consumer Price Index data) Economic Census
Services (except as separately specified below)	Census	Services Annual Survey Economic Census
Exceptions:		
Housing services (except rental value of farm dwellings)	Census	Housing Vacancy Survey Supplement to the Current Population Survey
Rental value of farm dwellings	Government or admin	U.S. Department of Agriculture
Electricity and natural gas	Government or admin	Energy Information Administration
Government hospitals	Government or admin; Census	Federal agency data Census of Governments
Motor vehicle leasing	Private/trade; Government or admin	R.L. Polk & Co. Bureau of Labor Statistics (consumer expenditures)
Parking fees and tolls	Government or admin	Federal Highway Administration
Railway transportation services	Private/trade	Amtrak
Intracity mass transit transportation services	Private/trade	American Public Transit Association
Taxicab transportation services	Census; Government or admin	Services Annual Survey Bureau of Labor Statistics (consumer expenditures)

Air transportation services	Government or admin	Bureau of Transportation Statistics
Food furnished to military	Government or admin	Federal budget data
Housing accommodation at schools	Government or admin	National Center for Education Statistics
Financial services furnished without payment	Government or admin	Federal agency data
Life, net household insurance, and net motor vehicles insurances services and workers compensation	Private/trade	A.M. Best
Medical care and hospitalization insurance	Government or admin; private/trade	National Center for Health Statistics A.M. Best
U.S. postal services	Government or admin	U.S. Postal Services
Higher education	Census; Government or admin	Census of Governments Annual Survey of State and Local Government Finances National Center for Education Statistics
Elementary and secondary schools	Government or admin	National Center for Education Statistics
Labor organization dues	Government or admin	Bureau of Labor Statistics Internal Revenue Service
Religious organizations' services to households	Private/trade	Giving USA
Net foreign travel	Government or admin	Bureau of Economic Analysis

Note: Data sources used for a given item can depend on whether an estimate is made for a benchmark year or a non-benchmark year.
Source: Bureau of Economic Analysis (2017b), Tables 5A and 5B

BEA documentation explains that PCE estimates "are based on statistical reports, primarily from the U.S. Census Bureau, but also from other government agencies; on administrative and regulatory agency reports; and on reports from private organizations, such as trade associations." (Bureau of Economic Analysis, 2017b, pp. 5-7). In Table 2.1, the three broad categories for data source are "Census," which includes data received from the U.S. Census Bureau, "Government or admin" which includes both survey and non-survey data received from other government agencies, and "Private/trade." In addition, in some cases an individual component in PCE may use multiple sources of information to derive an estimate. BEA uses estimating methods that adjust source data to the required NIPA concepts and that fill gaps in coverage and timing.

For the private sources, BEA undertakes efforts to ensure that the data are high quality by considering their accuracy, reliability, and relevance for the estimates. For example, BEA conducts computer edit checks for gross errors, identifies and analyzes outliers, and examines period-to-period changes. BEA also evaluates how representative the data are and how closely they fit with NIPA concepts. Federal sources of data can be expected to adhere to Information Quality Guidelines, which were described in this report's Introduction.

Documentation that Communicates Data Quality to Users. BEA clearly communicates the methodology and source data used in constructing the estimates of PCE as part of a comprehensive documentation of the NIPAs. The Transparent Reporting Project contacted selected users of BEA statistics to assess BEA documentation of PCE based on review of two documents: the chapter on PCE in BEA's comprehensive *Concepts and Methods of the U.S. National Income and Product Accounts* (Bureau of Economic Analysis, 2017b) and the 2017 article "Updated Summary of NIPA Methodologies" (Bureau of Economic Analysis, 2017c).

Assessing and Describing Data Quality. The BEA documentation states that PCE "provides a comprehensive measure of types of goods and services that are purchased by households. For example, it shows the portion of spending that is accounted for by discretionary items, such as motor vehicles, or the adjustments that consumers make to changes in prices, such as a sharp run-up in gasoline prices" (Bureau of Economic Analysis, 2017b, p. 5-1). Thus, the relevance of PCE goes beyond its contribution to measuring GDP, as important as that is. There can be interest in PCE or its own components that can be relevant for a government agency, a business, or a researcher. BEA documentation on NIPA begins by noting that, in part, "relevance refers to the ability of the accounts to provide summary and detailed estimates in analytical frameworks that help answer the questions being asked about the economy" (Bureau of Economic Analysis, 2017a, p. 1-8, emphasis added).

Disaggregating PCE beyond the broad distinction between goods (which are often subdivided between durable and nondurable) and services, PCE provides a "functional" classification composed of 13 more detailed categories such as "Clothing, footwear, and related services," "Transportation" and "Financial services and insurance" (Bureau of Economic Analysis, 2017b, p. 5-5). Statistics for the functional classification, along with major types of products, is available on-line (Bureau of Economic Analysis, 2019a). At an even finer level of product detail, detailed PCE information is available on-line for hundreds of types of products (Bureau of Economic Analysis, 2019b)

BEA integrates data from multiple sources to provide the granularity of product detail that users value. For example, in estimating the new motor vehicles component in PCE, BEA begins by acquiring detailed and comprehensive coverage on unit sales of autos and light trucks from *Ward's Automotive Reports* (Bureau of Economic Analysis, 2017b, p. 5-36). BEA then estimates the share of these sales that are purchased by persons using monthly data on new registrations from R.L. Polk & Co. Then these data are combined with average expenditure per transaction derived from monthly retail transaction prices by make, model, and trim level from J.D. Power and Associates to obtain an estimate of PCE for new motor vehicles. This approach allows BEA to produce timely estimates at a highly granular level of product detail.

BEA promotes access to statistical products, with documentation noting that users can access data on the BEA website (Bureau of Economic Analysis, 2017b, p. 5-7, footnote 7). Data can also be downloaded in various formats.

The dimension of punctuality refers to the time between a scheduled data release and when information became available. While not discussed in its methodology documentation, GDP is designated as a principal federal economic indicator. Statistical Policy Directive No. 3 emphasizes the importance of prompt release of such indicators and assigns responsibilities to publish a release schedule (50 Federal Register 38932-38934). Like other agencies responsible for principal federal economic indicators, BEA can be considered highly punctual in its development and release of GDP estimates.

To construct estimates of the NIPAs, GDP, and PCE, BEA adopts the conceptual framework laid out in the System of National Accounts (SNA), an international standard maintained and intermittently revised by the United Nations and others (European Commission et al., 2008). The SNA is the guiding framework that statistical agencies around the world follow, resulting in statistics that are coherent by being based on transparent and agreed-upon concepts, principles and approaches to combine and report information. In addition, GDP estimates from different countries achieve comparability when they are produced using the same framework. Following the revision to SNA in 2008, BEA developed a comprehensive update of the NIPAs, which included changes to PCE measurement, thus "improving consistency with international standards." (Bureau of Economic Analysis, 2017b, p. 5-4).

BEA documentation states that, "Accuracy may be described in terms of how close the estimates come to measuring the concepts that are designed to measure. In the case of GDP, the estimate is accurate when it captures all production for final use but does not include production for intermediate use" (Bureau of Economic Analysis, 2017a, p. 1-7, emphasis added). The distinction between intermediate and final expenditures is foundational for estimation of PCE and GDP to avoid double-counting economic activity. Because BEA estimates are a combination of survey-based, administrative and private data that are not collected within a common probability-based framework, BEA does not focus on accuracy in a statistical sense. Instead BEA relies on the concept of reliability—the repeated estimation of the same event—to assess the quality of its estimates. Reliability is discussed in more detail below when the revision process is described.

Users of NIPA estimates value the timeliness of the estimates, where timeliness refers to a small difference between the end of the period for which PCE or GDP estimates are prepared (e.g., a year or quarter) and the time when the estimate is released. The dimension of timeliness is treated as an aspect of relevance in some BEA documentation, which states that, in part, "relevance refers to the length of time before the estimates become available. Estimates that are not available soon enough for the intended use are not relevant" (Bureau of Economic Analysis, 2017b, 1-8, emphasis added). To address timeliness, BEA documentation notes that estimates are available for PCE and by major type of product (durable goods, nondurable goods, and services) each month (Bureau of Economic Analysis, 2017b, p. 5-6). Thus, by supplementing its annual and quarterly estimates with monthly statistics, BEA provides a very high level of detail in the temporal dimension. The advance quarterly estimates are available about one month after a quarter ends (Bureau of Economic Analysis, 2017c, p. 1).

BEA states that "there is an implicit tradeoff between timeliness and accuracy, so BEA has developed a release cycle for the estimates that addresses this tradeoff" (Bureau of Economic Analysis, 2017a, p. 1-8). The "release cycle" refers to how there is a sequence of estimates of GDP that cover the same time period. BEA produces successive vintages of so-called "current" estimates for quarter that are known as "advance," "second," and "third" estimates, which are available, respectively, about one, two, and three months after the end of the quarter (Fixler, et al., 2018). Following the set of current estimates, subsequent revisions are made that constitute successive "annual update" estimates and one or more "comprehensive update" estimates.

Current quarterly estimates of PCE and the other NIPA estimates are available most quickly. However, they use partial and preliminary data that are not as comprehensive as the annual source data, which impinges on accuracy Some detailed components of the current estimates of PCE are prepared through extrapolation or interpolation using data that are relatively frequent and timely to serve as indicators of the co-movements of the series over time. Documentation describes how the accuracy of extrapolators are improved through weighting, filling in gaps in coverage, bias adjustments, averaging with other extrapolators, and benchmarking and balancing.

Annual updates also include indicators, but they are less important than for current quarterly estimates. At five-year intervals the Census Bureau conducts an economic census and BEA constructs its NIPA estimates for what BEA calls a "benchmark" year—a year that provides the most comprehensive source data for NIPA. Table 2.2 illustrates the shares of the major types of source data used in estimating PCE for the first annual update and for the benchmark year.

Table 2.2. Expenditure shares within Personal Consumption Expenditures (2016)

Type of data	Expenditure shares in first annual update (percent)	Expenditure shares in benchmark year (percent)
U.S. Census Bureau data	78	85
Government and administrative sources	11	8
Private and trade data	8	4
Multiple data sources	3	3

Note: Expenditure shares exclude non-profit institutions serving households.
Source: author's calculation based on the tables in Bureau of Economic Analysis (2017b) and Table 2.4.5U in the Underlying Detail section of the GDP Interactive tables.

The share of data acquired from the U.S. Census Bureau increases from 78 percent for the first annual update to 85 percent for the benchmark year. This increase mainly reflects the incorporation of the economic census data for some estimates which become available every 5 years.

In addition, the NIPAs are regularly updated to reflect changes in economic concepts and methods necessary for the accounts to provide a relevant and accurate picture of the evolving U.S. economy. These updates range from expanding the definition of investment to include research and development activity to updating seasonal adjustment factors to reflect the most recent seasonal patterns.

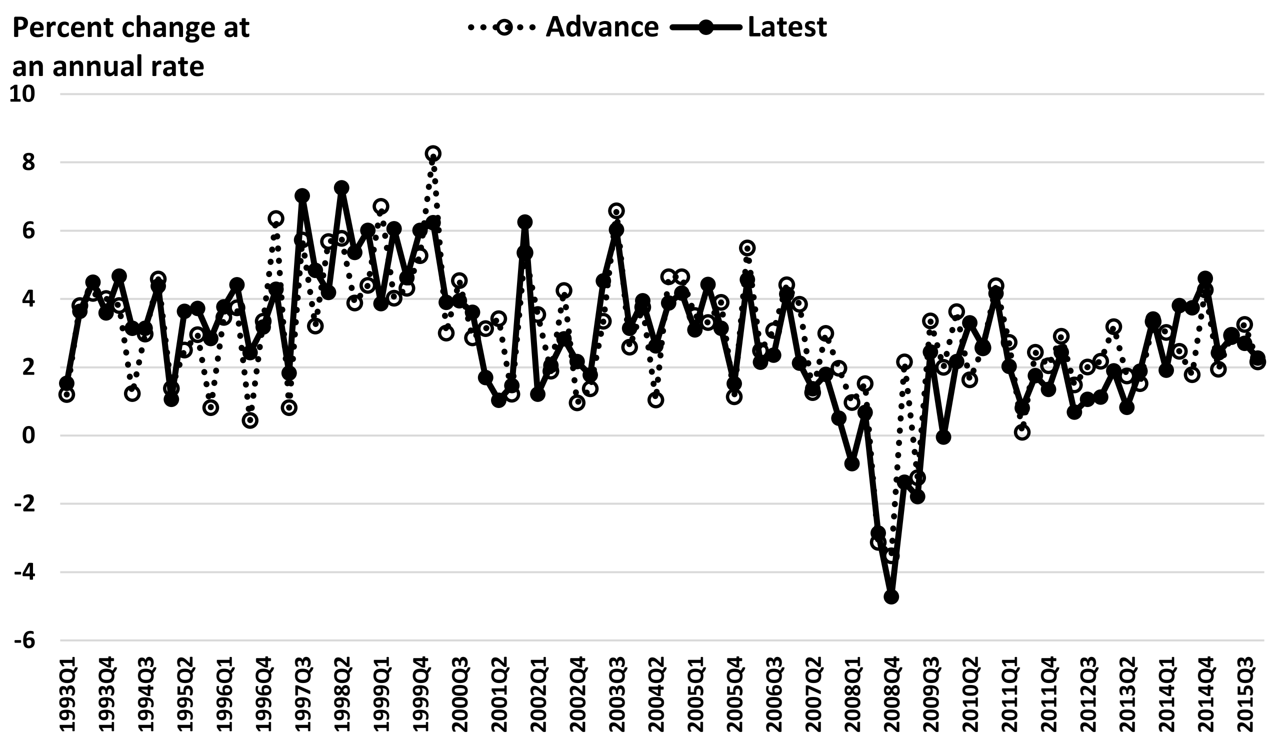
As revisions occur, BEA provides descriptions of planned update procedures in publications that examine methodology and source data. Updated estimates are released in Survey of Current Business articles and on the BEA website. BEA news releases routinely place a revision in the context of recent revisions and compare the revision with overall average revisions. In news releases and in annual and comprehensive update articles, a section of technical notes section discusses the assumptions and methods.

The release cycle and its periodic revisions to GDP estimates (for a given time period) lead to concept of reliability. In statistical terms, reliability concerns the repeated estimation of an event. At BEA, "Reliability refers to the size and frequency of revision to the NIPA estimates. An important indicator of reliability is the effectiveness of the initial estimates of GDP in providing a useful picture of U.S. economic activity" (Bureau of Economic Analysis, 2017a, p. 1-8, emphasis added). The reliability of the estimates is assessed by examining the performance of the successive vintage estimates; for example, an examination of the closeness of estimates as they are revised when more complete data become available.

There are periodic studies of GDP revisions that study the reliability of the estimates over the release cycle and examine evidence for any systematic overstatement or understatement. BEA's principal standard of reliability is to examine the revisions from its early estimates to its "latest" estimates, most of which have been through at least one comprehensive update; the "latest" estimates are deemed to be the most reliable because the comprehensive update incorporates all the available source data that are believed to be the most reliable. The periodic studies have "confirmed that the initial estimates provide a reliable indication of whether economic growth is positive or negative, whether growth is accelerating or decelerating, whether growth is high or low relative to trend, and where the economy is in relation to the business cycle" (Bureau of Economic Analysis, 2017a, p. 1-8). Such studies are an important vehicle for BEA to ensure the quality of the estimates. Users of BEA data can use the findings of these studies to better understand the nature of the BEA estimates.

The most current of these periodic studies that examine the reliability of GDP estimates is Fixler, et al. (2018). According to the study's results, both the pattern and magnitude of the revisions indicate that the early estimates are reliable. Figure 2.1 illustrates this finding using 1993-2015 data for two vintages of estimates of real (inflation-adjusted) PCE—the advance estimates and the latest estimates.

Figure 2.1 Real personal consumption expenditures, for two vintages of estimates, 1993 – 2015



Source: authors' calculation using the database that underlies tables in Fixler, et al. (2018)

The chart shows that, with few exceptions (such as the sharp divergence in the growth rate in 2009 Q1 during the depths of the financial crisis), the overall pattern of movement in PCE is little changed by revisions. Similarly, for GDP in general, the early estimates provide an accurate general picture of economic activity. That is, revisions to GDP estimates do not substantially change BEA's measures of long-term growth, the picture of business cycles, and the trends in major components. Policymakers can use these GDP estimates as reliable measures of economic activity.

Costs of Documentation. BEA spends about $2 million annually to prepare documentation for all BEA programs, inclusive of the PCE component of GDP.

Responses to the Customer Survey. The agencies in the Transparent Reporting Project solicited participation in the customer survey from two groups—informed consumers and researchers. BEA used the project as an opportunity to focus exclusively on informed consumers, inviting 6 to participate in the survey of whom 1 responded. Despite the low response for PCE documentation, there are still benefits in considering the user assessment both for PCE and in combination with the other products selected for this report. The detailed responses to the Likert items and other structured items for the PCE are provided in Appendix 5. An item on the survey may not have been applicable for assessment of BEA documentation if, for example, the item referred to linking surveys to administrative data. The free responses for assessment of PCE documentation are highlighted here.

One item posed by the customer survey asked, “How did you use the data documentation or other information provided by the agency to inform or address your information needs?” (#19 in Appendices 4 or 5; A19 on the informed consumer instrument in Appendix 2). The response mentioned methodology papers and the “truly outstanding” section of Frequently Asked Questions (FAQ) on the BEA website. Other agencies that do not have a FAQ section on the website could consider developing such a resource. A FAQ section can enable users to access information that may not be contained in the details of documentation or that might be difficult to locate.

Another item was “What information on data sources used in {Product} did you find useful?” (#26; B4) The response was that it was helpful to know which components are based on unpublished data.

While most of the customer survey involved assessment of current documentation, one of the items asked users about improvements: “What suggestions do you have for improving documentation about source selections made for {Product}?” (#27; B5). The suggestion was to provide an online data dictionary and links to websites that provide some of the raw data when data sources are not statistical agencies such as BLS or the Census Bureau. Other agencies could also consider developing an online glossary for key terms.

References to Chapter 2 (National Accounts)

Council of Economic Advisers. 2019. *Economic Report of the President, 2019*. https://www.whitehouse.gov/wp-content/uploads/2019/03/ERP-2019.pdf

European Commission, International Monetary Funds, Organisation for Economic Co-operation and Development, United Nations, and World Bank. 2009. *System of National Accounts, 2008*. New York https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf

Federal Register. 1985. Volume 50, No. 186. 38932-38934. September 25, 1985. https://www.bea.gov/sites/default/files/2018-05/federalregister09251985.pdf

Fixler, Dennis, Danit Kanal and Pao-Lin Tien. 2018. "The Revisions to GDP, GDI, and their Major Components." *Survey of Current Business*, Vol. 98, No. 1, January 2018. https://apps.bea.gov/scb/toc/0118cont.htm

U.S. Bureau of Economic Analysis. 2000. "GDP: One of the Great Inventions of the 20th Century." *Survey of Current Business*. Vol. 80, No. 1. January 2000. https://apps.bea.gov/scb/pdf/BEAWIDE/2000/0100od.pdf

U.S. Bureau of Economic Analysis. 2017a. "Chapter 1. Introduction," in *Concepts and Methods of the U.S. National Income and Product Accounts*. Washington, DC: BEA, November 2017. https://www.bea.gov/resources/methodologies/nipa-handbook

U.S. Bureau of Economic Analysis. 2017b. "Chapter 5. Personal Consumption Expenditures," in *Concepts and Methods of the U.S. National Income and Product Accounts*. Washington, DC: BEA, November 2017. https://www.bea.gov/resources/methodologies/nipa-handbook

U.S. Bureau of Economic Analysis. 2017c. "Updated Summary of NIPA Methodologies," *Survey of Current Business* 97, November 2017. https://apps.bea.gov/scb/toc/1117cont.htm

U.S. Bureau of Economic Analysis. 2018. "Consumer Spending," webpage updated September 6, 2018. https://www.bea.gov/resources/learning-center/what-to-know-consumer-spending

U.S. Bureau of Economic Analysis. 2019a. Table 2.3.5.U. "Personal Consumption Expenditures by Major Type of Product and Major Function," webpage updated April 26, 2019. https://apps.bea.gov/iTable/iTable.cfm?ReqID=19&step=2#reqid=19&step=2&isuri=1&1921=underlying

U.S. Bureau of Economic Analysis. 2019b. Table 2.4.5.U. "Personal Consumption Expenditures by Type of Product," webpage updated April 26, 2019. https://apps.bea.gov/iTable/iTable.cfm?ReqID=19&step=2#reqid=19&step=2&isuri=1&1921=underlying

3. Integration of Administrative Data and Survey Data

— National Postsecondary Student Aid Study (NPSAS) —
National Center for Education Statistics (NCES)

— Linkage between the National Hospital Care Survey (NHCS)
and the National Death Index (NDI) —
National Center for Health Statistics (NCHS)

— The Veteran Population Projection Model 2016 (VetPop2016) —
National Center for Veterans Analysis and Statistics (NCVAS)

In the Transparency Reporting Project, three statistical products make use of both administrative data and survey data: the National Postsecondary Student Aid Study (NPSAS); the linkage between the National Hospital Care Survey (NHCS) data and the National Death Index (NDI), and the Veteran Population Projection Model 2016 (VetPop2016). A common practice used to produce all three products is linking survey and administrative record data through matching at the individual person or institutional respondent level. For all three products, such linking is conducted for statistical purposes only.

The National Center for Education Statistics (NCES) collects data for NPSAS, which provides information on how students pay for postsecondary education. The most recent NPSAS estimates, for 2015-16, find that nearly three-quarters (about 72 percent) of undergraduates received some type of financial aid (grants, loans, work-study, and other) which averaged $12,300 among those receiving aid (Radwin et al., 2018).

The National Center for Health Statistics (NCHS) designed the NHCS to "provide accurate and reliable health care statistics that answer key questions of interest to health care and public health professionals, researchers, and health care policy makers" (National Center for Health Statistics, 2019). The linkage of the NHCS data to the NDI expands the analytic utility of the NHCS because the NDI provides information on factand cause of death; a limitation of the NHCS and the linked data is that they are not nationally representative. One example using the linked data noted that among patients who had an inpatient or emergency department encounter in 2014 and were diagnosed with Alzheimer Disease (the most common form of dementia), 44 percent died in 2014 or 2015, with causes of death that included diseases of the circulatory system, nervous system, and mental, behavioral, and neurodevelopmental disorders (Bercovitz et al., 2018, p. 4).

The National Center for Veterans Analysis and Statistics (NCVAS) developed the Veteran Population Project Model 2016 (VetPop2016). Using VetPop2016, NCVAS prepares an official veteran population projection for VA. VA also uses results from the model to project medical and financial resource needs down to county-level VA service areas. Based on the model, the total veteran population is projected to decline over the next twenty years from about 20.0 million in 2017 to 13.6 million (National Center for Veterans Analysis and Statistics, no date). Beginning in 2016, veterans of the Gulf War Era became the largest cohort of veterans, outnumbering those of the Vietnam Era.

This chapter highlights data documentation and quality issues for each of these three products, including users' assessments of that quality. While the three case studies all use survey and administrative data, the reviews below emphasize different aspects of the issues for the three statistical products under study. The NPSAS case study provides more discussion than the other two issues concerning coherence and comparability. The case study on linkage between NHCS data and NDI provides relatively more details about linking issues and the possible types of errors that emerge from the linkage process itself—errors that are not attributable to either data source independently. The case study on the VetPop2016 model gives more details on how data quality is defined and described in a setting that relies heavily on administrative data. As a set, these three case studies provide an overview on transparent reporting for statistical products that combine survey and administrative data.

3(a). NPSAS

The Statistical Product and its Data Sources. NCES, within the U.S. Department of Education's Institute of Education Sciences, is the primary federal entity for collecting and analyzing data about education in the United States and other countries. Part of the agency's mission is to provide information about how students and their families pay for postsecondary education. NPSAS was developed by NCES to meet this information need. NPSAS is a nationally representative study of students attending postsecondary institutions in the United States and is traditionally conducted every three or four years. The studies draw on data from student interviews, data provided directly from postsecondary institutions, and six distinct administrative sources.

NPSAS provides detailed measures on student financial aid and borrowing as well as other key indicators of postsecondary education (e.g., enrollment) and demographics. While the student interview has traditionally played a significant role in creating the study's measures, the increasing quality and availability of administrative data has facilitated a shift over time. Now, many constructs rely less on self-reported information and more exclusively on administrative records. In fact, improvement in source data made possible an NPSAS collection in 2018 that relied exclusively on administrative data.

As an integrated data product, NPSAS draws upon multiple data sources, some of which are described in Table 3.1.

Table 3.1. Student-level administrative Sources in NPSAS

Source	Description and Use
The Central Processing System (CPS)	To be eligible for federal financial aid, students must file the Free Application for Federal Student Aid (FAFSA). These data contain financial and demographic information on students and their families but is limited to those who apply for federal financial aid.
The National Student Loan Data System (NSLDS)	This federal database tracks Pell grant and federal student loan award amounts and disbursements. NPSAS also uses information housed in NSLDS related to postsecondary enrollment, loan repayment, income, and demographic information. Data are only available for federal student loan and Pell grant recipients.
The Veterans Benefits Administration (VBA)	These data contain information related to Veterans and Active Duty service members of the US Military. Used primarily to identify military sample members and to aggregate federal veterans' education benefit payments made during the study time-frame, data are only available for VA beneficiaries.
The National Student Clearing House (NSC)	These data primarily measure students' enrollment at all postsecondary institutions that participate in or use NSC's services. NPSAS matches sample members to this database and use the data to derive enrollment, degree, and major/field of study information.
Testing Agencies: The College Board (SAT) and ACT	NPSAS matches to both College Board and ACT to obtain standardized test scores. These data also include information on high school course-taking and are available for students who took either exam.

This list represents the data that are collected at the student level. NPSAS draws on other data sources as well, notably NCES' Integrated Postsecondary Data System (IPEDS) which is used to create variables that describe the characteristics of institutions that students attend and occasionally used for logical imputation in NPSAS. In addition to these administrative data, NPSAS also collects data directly from postsecondary institutions (henceforth Student Records).

Documentation that Communicates Data Quality to Users. For the Transparent Reporting Project, selected users of NPSAS were sent the customer survey along with five pieces of NPSAS documentation. One was the most current detailed report, the NPSAS Data File Documentation for 2015-16 (Wine et al., 2018). A second was the Full-scale Methodology Report (Cominole et al., 2010), a background document containing a major review of methodology. Users were also sent three documents derived from the NCES website: "About NPSAS," "State Oversamples," and "Data Information for Postsecondary Sample Surveys" (National Center for Education Statistics, no date a, b, c)

Assessing and Describing Quality. Multiple sources of data are often blended together to form "derived" or "composite" variables. For example, the months during which students were enrolled in college incorporate self-reported interview responses, Student Records, NSC, and NSLDS data. Each of these data sources, on its own, contains missing enrollment information, which generates challenges for data quality when combining information on enrollment from multiple sources. Some variables draw on only one source of data because of the completeness and accuracy of the source. For example, federal student loan amounts are drawn only from NSLDS. NSLDS is heavily audited to ensure it is properly tracking loans and their repayment through time. Other sources of information about student loans exist including records held by universities and the students themselves. Using information from these sources would not improve the accuracy of the data (and reduce it in terms of student self-reports in the short NPSAS interviews) and increase the overall response burden of NPSAS.

The process of developing derived variables results, in the first instance, in missing values for many students.. The derived variables often draw information from multiple sources of data for the student and these sources are not consistently complete across all students sampled for NPSAS. To improve the accuracy and reliability of information generated from NPSAS, missing values within derived variables are stochastically imputed with the goal of achieving nationally representative estimates from the resulting data.

Work undertaken to improve accuracy and reliability creates trade-offs between those dimensions and timeliness The sheer volume of derived variables and the need to ensure that all imputations produce internally consistent information requires an extensive set of logic and range checks against other information available for each student in NPSAS. This work prolongs the time it takes from the end of data collection to the release of the final data to the public.

To maximize accessibility, the finished data are released in two forms. The first form is a public version of the data that can be accessed at the NCES-sponsored website, DataLab (National Center for Education Statistics, no date d). DataLab includes three interfaces that can be used to create and view estimates of NPSAS variables: QuickStats generates simple summary statistics; PowerStats enables users to create more advanced cross-tabulations and linear and logistic regressions; and TrendStats enables users to compare variables across multiple administrations of NPSAS, as far back as 1996. Embedded in these systems is a tool by which the agency tracks the usage of individual variables on PowerStats, both at the univariate and bivariate (combinations of variables) level. NCES uses information from the tool to gauge variable relevance to users, which informs future NPSAS data collections. For example, variables that are seldom or never used are evaluated for potential removal from the next NPSAS in combination with information on item quality and the need for the item to address a specific critical question. Data are safeguarded through several DataLab features, including prohibiting estimates from being generated for relatively small subpopulations, preventing direct access to the data that underlie DataLab processing, and perturbing the data before they are entered into DataLab. While these features maintain the confidentiality of the data, they can impinge on data quality, such as the granularity of the available estimates.

The second form of data released is a restricted-use file (RUF), which requires users to apply for a license through NCES. Applications must include a plan to maintain data security per NCES standards. These data are far more extensive than the public-facing versions in DataLab. They give advanced researchers with approval the added flexibility to analyze sophisticated statistical models and procedures to generate their own estimates. Data are safeguarded through a number of protections, including audits that ensure adherence to security protections, output review by NCES, and data perturbations before researchers access the files.

Due to the complexity of NPSAS, providing information on how its data are created is essential to end users of all skill and interest levels. To promote access and clarity an array of six distinct products that describe the data are provided, each with a specific type of user in mind. These multiple vehicles for documentation also ensure rules are followed on disclosure of information. Three documentation products are accessible to all NPSAS users: a formal and extensive Data File Documentation (DFD) report; codebooks for DataLab variables that provide detailed information about characteristics of the variables; and a “Data Info/Issues” page on the NPSAS website. Users with RUF access receive three additional pieces of documentation: a README file explaining how to correctly read the data; codebooks for all derived and source file datasets; and SAS code that created the derived variables.

The DFD for a typical administration of NPSAS is over 200 pages long and includes another 1,000 pages of appendices (Wine et al., 2018). This document describes all phases of the data collection process to inform those accessing the data about its accuracy. The DFD also serves to quantify the burden the study imposes on the students and institutions who provide data; the burden is measured in the aggregate and for survey items. Chapter 5 of the DFD describes “Administrative Records Matching Overview and Outcomes.” The chapter includes details on the matching procedures and rates, disaggregated by selected sampling-frame characteristics, such as institution and student type. The DFD also provides details on sampling and weighting which are designed to produce high quality estimates.

Codebooks are also provided to the public as part of PowerStats. Figure 3.1 illustrates the codebook documentation associated with the variable, TITIVAMT (Total amount of federal Title IV financial aid received by a student) during the 2015-16 academic year.

[Table follows on next page]

Figure 3.1. Public Codebook entry for the NPSAS:16 variable TITIVAMT

Subject: Financial aid: Federal
Label: Total federal Title IV aid
Name: TITIVAMT
Description: Total amount of federal Title IV financial aid received during the 2015-16 academic year.
Source: NSLDS:16, NPSAS:16 Student Records, NPSAS:16 Interview
Descriptive Statistics:

Value	Percentage	Label
Continuous	54.42	Positive values, see statistics below
0	45.58	{Zero}

Minimum	Maximum	Average	Standard Deviation
1.00	77403.00	8572.27	6865.12

Weight used in frequency: (WTA000)
Notes: Equal to the sum of Title IV loans (including Parent PLUS loans) (T4LNAMT2), Pell grants (PELLAMT), Supplemental Education Opportunity Grants (SEOGAMT), TEACH Grants (TEACHGRT), and federal college work-study (TFEDWRK). Unlike total federal aid (TFEDAID), this variable excludes non-Title IV programs like federal health professions loans and other federal grant programs (OTHFDGRT). TEACH grants (TEACHGRT) are included in TITIVAMT since they are also disbursed under Title IV. NPSAS:12 included TEACH grants as other federal grants (OTHFDGRT) and did not report them separately. Also see TFEDAID2, which includes federal Veterans' benefits and Department of Defense grants (VADODAMT).
Code: TITIVAMT = SUM(OF T4NAMT2 PALLAMT SEOGAMT TEACHGRT TFEDWRK);
Applies to: All respondents.

In addition to basic fields (such as variable name (Name), variable label (Label), and variable description (Description)), the codebooks include other key pieces of metadata for public users. First, the Source field lists all the data sources that were used to create the variable, listed in terms of importance to thatvariable. The Notes field is used to expand upon the description and source prioritization, particularly when variables are complex and use more than one source of data. Finally, the Code field expresses pseudo-SAS code that can be used to re-create the variable and to state programmatically how the variable was created using other variables available on the dataset. In the case of TITIVAMT, users can read of its summative nature in the Notes field and know precisely which other variables are contained within it based on the Code field.

A separate codebook is provided to RUF users. This codebook includes the fields provided on the public PowerStats, but also provides unweighted counts and percentages. In addition to the variables available on PowerStats, the RUF datasets and codebooks also provide what are known as "z-variables" – variables that indicate the number and percentages of individuals in the sample that receive each source, inclusive of logical imputation (e.g., zero amounts for non-borrowers) and those who had the corresponding variable stochastically imputed.

The README file is another product only available to RUF users and is included with the data files. The NPSAS:16 RUF contains two derived variable files (one for undergraduate students and another for graduate students), a weight file, and 11 source files. Additionally, it contains 92 SAS programs that create the dataset and prepare the imputations; a handful of reports; copies of forms (e.g., the Free Application for Federal Student Aid (FAFSA)); programs to unpack the data in SAS, Stata, and SPSS formats; and other items that NCES provides for users (e.g., updates to previous NPSAS administrations). While it contains information that is also provided on the Data Info/Issues page and DFD, the unique feature of the README is that it aids users in navigating the voluminous information provided, thereby improving the clarity of the data themselves and the processes to follow to use the data properly

Since its inception, NPSAS has been conducted nine times. Over time, the study has expanded to include new administrative sources, which influences both the data themselves and how they are documented. The addition of administrative sources is generally thought to increase accuracy. At the same time, though, for mature studies such as NPSAS, such a change can disrupt the coherence and comparability of the same construct over time. For example, NPSAS:16 was the first NPSAS administration to use data from the Veterans Benefits Administration (VBA). Inclusion of the VBA data increased the accuracy of the amounts of veterans' education benefit in the data. However, this change resulted in a significant positive deviation from the trend line for these benefits; NPSAS studies prior to NPSAS:16 had used student self-reported amounts from the interview and Student Records, where possible. The effect on the data was so dramatic—the average benefit amount doubled between 2012 and 2016—that it required the creation of specialized variables to sustain retrospective comparison of veterans' education benefits. It also required communicating to end users the shift in data sources and its effects in the codebook's Notes field. Users were also notified through the NPSAS Data Info/Issues web page and the README file.

While efforts are made to document the data creation process and use of administrative sources, there are constraints on the level of detail that can be provided. First, as indicated above in Table 3.1, NPSAS uses proprietary data from NSC, ACT, and the College Board. These private organizations require data use agreements that prohibit releasing the source files and detailed documentation. A similar arrangement with VBA exists that precludes the release and documentation of the source data. Users are unable to completely recreate variables that incorporate these sources, so the documentation does not include complete replicability. A second limitation on the detail of documentation involves the underlying structure of federal databases such as NSLDS, which could pose security risks for these databases that primarily serve to manage student loans. However, despite these limitations, documentation does provide users with details about data sources and the logic used to construct the variables. In so doing, the documentation enables end users to evaluate fitness for their own uses to the greatest extent possible.

Costs of Documentation. The cost of documenting the integration processes and documenting how to use the integrated data are relatively inexpensive in relation to the cost of the actual integration processes themselves. Overall, the most detailed source on data integration in NPSAS is the "National Postsecondary Student Aid Study (NPSAS) Data File Documentation." A new report is produced with each collection. The most recently released report in the series is Wine et al. (2018). The full report cost $300,000. Approximately 20 percent of the report considered issues of data integration, including chapters on how the institutional and student collections were designed, how administrative data were matched to survey information, and related appendixes. Thus, the costs of documentation specifically tied to NPSAS as an integrated data product is about $60,000.

Responses from the Customer Survey. Overall, evaluations of the NPSAS documentation were provided by 13 users of the data. Generally, researchers who evaluated the documentation for NPSAS thought the documentation was informative, clear, and useful. Most of the questions had 5-category response scales with 1 meaning Very Dissatisfied and 5 meaning Very Satisfied. Of the 42 items on this scale, 5 had average scores below 4 (ranging from 3.2 to 3.9) and the rest were at 4 or above (up to 4.9). Researchers like the clarity and detail of the documentation overall, and the descriptions of how data integration was undertaken from selection of source information to integration procedures themselves (scores ranged from 4.5 to 4.7 across these topics). Researchers were less satisfied with timeliness and punctuality of data release overall, how disclosure prevention and new data sources might affect estimates in general, and detailed results of integration procedure evaluation (scores ranged from 3.2 to 3.9 on these topics).

The items that allowed users to provide unstructured, free responses provided additional information. The item "How did you use the data documentation or other information provided by the agency to inform or address your information needs?" (#19 in Appendices 4 or 5; A19 on the researcher instrument in Appendix 3) elicited a variety of responses. Based on the documentation, users were able to review data reliability, sampling procedures, methods for calculating variables and which students were included, and correct interpretation of the data. One user identified a need that researchers have for agency-produced tables to accompany a micro-level dataset: "I typically try and run tables that match with a published table to make sure I am doing it the same way. Documentation helps me with this." Another user noted how helpful the staff were in explaining variable definitions, data sources, and how to generalize from the sample.

In response to the item "What information on data sources used in {Product} did you find useful?" (#26; B7), users wrote about the importance of knowing where the data came from, such as institutional records, federal databases, or student interviews. A comparison of the 2016 version of NPSAS to previous studies was helpful. The usefulness of quality measures was also mentioned.

It can be useful to distinguish between the availability of information and the ease with which the information is located. Some suggestions for improvements addressed the latter. A suggestion for improving documentation about source selection (#27; B8) noted that it was difficult for the user to navigate the website to obtain the needed information, which is a different concern than the absence of the needed information. Similarly, when asked for suggestions on how to improve documentation about procedures used to integrate information from different sources (#32; C11), a user reported that locating suitable data took only a short amount of time but locating the needed methodological notes took much longer; the user suggested additions to the documentation on methodological steps.

To improve documentation on how to properly use the product (#46; D12), one user suggested simplifying some of the information on sampling weights provided in the PowerStats section of DataLab. This suggestion points to the challenge of writing materials that are both brief and clear and the challenge of serving a wide variety of users, some of whom may prefer only the fundamental information and others of whom want more detail.

The customer survey included an item on improving data quality: "In contrast to improving documentation, do you have suggestions for improving the statistical product itself or its quality?" (#65; F1). Some users provided suggestions, while others expressed appreciation for the current form of NPSAS. One user suggested additional items on working while in school, and on student financial wellness and basic needs. After stating that the addition of the administrative data to the product enabled it to be timelier and representative at the state level, one user suggested including additional administrative data. One user gave no suggestions for improvements, offering that "NPSAS is superbly conducted."

3(b). NHCS-NDI Linkage

The Statistical Product and its Data Sources. The methodological documentation for the linkage between the National Hospital Care Survey (NHCS) data and the National Death Index (NDI) begins with a broad overview of the agency and the data sources: "As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. In addition to collecting and disseminating the Nation's official vital statistics, NCHS conducts several population-based surveys and health care facility establishment surveys, including the National Hospital Care Survey (NHCS)" (National Center for Health Statistics, 2018a, p. 3)

Historically, NCHS has conducted three national surveys annually across five ambulatory and hospital-based settings: physician offices, inpatient settings, emergency departments (EDs), outpatient departments (OPDs), and hospital ambulatory surgery locations. In an effort to streamline data collection across health care settings and move toward collecting health care utilization data by electronic means, NCHS launched the NHCS, which integrates the National Hospital Discharge Survey (NHDS), the National Hospital Ambulatory Medical Care Survey (NHAMCS) and the collection of substance-use related ED visits.

Before NHCS' implementation, the NHDS, conducted by NCHS during 1965–2010, provided critical information on the utilization of the nation's nonfederal short-stay hospitals and on the nature and treatment of illness among the hospitalized population. NHAMCS, also conducted by NCHS, has provided data annually since 1992 about the nation's use of EDs and OPD visits, and since 2009, on the use of hospital ambulatory surgery locations. These data have been extensively used for monitoring changes and analyzing the types of ambulatory care provided in the nation's hospitals.

The goal of NHCS is to provide timely and reliable health care data for hospital-based utilization. To accomplish this goal, NHCS has five objectives. First, NHCS is moving toward all electronic data collection, particularly using electronic health record data as it becomes more widely available. Second, when the survey is fully implemented, NHCS will provide nationally representative utilization statistics for hospital inpatient care, ambulatory medical care, and ambulatory surgery from a national probability sample of hospitals. Third, NHCS data will permit special studies to be conducted for both inpatient and ambulatory care as policy and research needs arise. Fourth, with the collection of personally identifiable information (PII) (e.g., name, address, and social security number), NHCS data can be linked across hospital settings within a sampled hospital and to outside data sources, such as the NDI or data from the Centers for Medicare and Medicaid Services (CMS). Finally, when fully implemented, NHCS will produce nonidentifiable microdata public-use files of inpatient discharges and ED and OPD visits, including ambulatory surgery, and will disseminate timely data that can be used by health policy researchers, the public, and the research community. Using these data files, researchers will be able to study trends and changes in health care practices and changes in patterns of health care-seeking behavior.

The target universe of NHCS is inpatient discharges, also called inpatient hospitalizations, and in-person visits made to EDs and OPDs, including ambulatory surgery, in noninstitutional nonfederal hospitals in the 50 states and the District of Columbia that have six or more staffed inpatient beds. No geographic primary sampling units are used in this design, and there are no certainty hospitals (hospitals with a 100 percent selection probability). The 2014 NHCS sample was developed from the 2010 spring release of "Healthcare Market Index" and "Hospital Market Profiling Solution, Second Quarter, 2010," both by Verispan and will be updated periodically for future surveys. The 2014 NHCS sample consists of 581 hospitals: 506 acute care hospitals and 75 other specialty hospitals, including children's, psychiatric, long-term acute care, and rehabilitation hospitals.

For the 2014 data collection, 94 hospitals out of the 581-hospital sample provided inpatient claims data, and 88 of the 94 hospitals that provided inpatient data also provided ambulatory claims data (a response rate of 16.2 percent and 15.1 percent, respectively). Participating hospitals were asked to provide all encounters in inpatient and ambulatory settings in the 2014 calendar year. The unweighted total number of encounters was approximately 1.7 million inpatient discharges, or inpatient hospitalizations, (1.5 million non-newborn inpatient discharges), and 4.5 million ED visits.

The NDI is a centralized database of United States death record information on file in state vital statistics offices. Working with these state offices, NCHS established the NDI as a resource to aid epidemiologists and other health and medical investigators with their mortality ascertainment activities. (National Center for Health Statistics, 2017). The NDI contains person-level information on date and causes of death collected from state death records.

Through its data linkage program, NCHS has expanded the analytic utility of the NHCS by augmenting it with mortality data from the NDI. The resulting linked 2014 NHCS- 2014/2015 NDI linked data file provides the opportunity to examine the incidence and cause of death among participating hospitals' inpatient and emergency department patients. The linked NHCS-NDI data support analysis of health outcomes that are not attainable using the survey data alone.

Documentation that Communicates Data Quality to Users. The users who received the Transparent Reporting Project's customer survey also received three documents on the 2014 NHCS- 2014/2015 linked NDI data: the methodology report (National Center for Health Statistics, 2018a) and two codebooks (National Center for Health Statistics, 2018b and 2018c). It should be noted that the linked NHCS-NDI data are only available as restricted-use files accessible through the NCHS Research Data Center (National Center for Health Statistics, 2019a).

Information regarding the NHCS data is provided in data file dictionaries published on the NHCS website (National Center for Health Statistics, 2019).

Assessing and Describing Quality. The matching methodology and analytic considerations linkage report communicates the relevance of the NHCS-NDI linkage by explaining the purposes of the NHCS and of its linkage to the NDI in the report's introductory and background sections (National Center for Health Statistics, 2018a, pp. 3-4).

For integrated data, issues of accuracy involve, in part, the representativeness of both sources of data. As indicated in the methodology report, the NDI data originate as state-level administrative data. Such data are deemed to be highly accurate in the sense of being a comprehensive set of all death certificates recorded by state vital statistics offices. The NDI contains death certificate information for death records on file in state vital statistics offices for all 50 states, the District of Columbia, New York City, Puerto Rico, and U.S. Virgin Islands. Deaths that occur outside of these U.S. registration areas are not included in the NDI. The methodology report also noted that the 2014 NHCS was not nationally representative (National Center for Health Statistics, 2018a, p. 3), which affects the quality of results based either on the 2014 NHCS or the linked 2014 NHCS-2014/2015 NDI data file. At the same time, the methodology report for the linked 2014 NHCS-2014/2015 NDI data file provides researchers with important information on the potential bias when using non-nationally representative survey data. NCHS offers guidance on how to mitigate this potential bias, by suggesting researchers consider controlling for hospital characteristics when conducting statistical analyses. (National Center for Health Statistics, 2018a, p. 7).

For a linked data file, the linkage process itself has implications for accuracy. Even though NHCS is an establishment survey for which hospitals are the sampling unit, it collects PII from patient encounter records, such as name and date of birth, which are used only for statistical purposes. The PII makes possible the linkage between the NHCS and the NDI as well as linkages of episodes of care across hospital units. (National Center for Health Statistics, 2018a, p. 4). However, some of the participating hospitals did not include the PII data required for linkages for most or all of their records and so therefore their patient populations were considered ineligible for linkage. (National Center for Health Statistics, 2018a, p. 6). NCHS is also only able to conduct linkage activities for NHCS patient records that meet the minimum threshold for completeness of PII. Patient records that do not meet these criteria are excluded from linkage and are flagged with a variable to indicate that record is linkage ineligible. Analysts are provided with guidance on how to exclude these records when conducting mortality analysis and how to review the linkage eligibility status within NHCS participating hospitals.

A section on linkage methodology and a five-page appendix provides detail on the linkage process, which has implications for the accuracy of the linked file. In any linkage or matching methodology, it is possible to generate false positive and false negative results. A false positive (a false match) occurs when the matching process mistakenly concludes that a record in the NHCS and a record in the NDI belong to the same person when, in fact, the patient in the NHCS and the deceased person in the NDI are different people. A false negative (a miss) occurs when no match is found for a given NHCS record even though that person's death information is included in the NDI. Standard linkage processes attempt to keep both false positives and false negatives to a minimum. NCHS provides users with detailed information on the linkage process, including variables used for linking, matching algorithms, probabilistic matching thresholds, and criteria for assigning vital status. (National Center for Health Statistics, 2018a, p. 11) As noted in the report the thresholds for the linked 2014 NHCS-2014/2015 NDI were set at the level that produced the lowest estimated total error (Type I and Type II). (National Center for Health Statistics, 2018a, p. 15)

The linkages between the NHCS records and NDI records were based on both deterministic and probabilistic approaches. The probabilistic approach performs weighting and link adjudication following the Fellegi-Sunter method, which is the foundational methodology used for record linkage. This linkage methodology estimates the likeliness that each pair of records is a match before selecting the most probable match above a defined threshold. The linkage process generates a "probability of match validity" for each candidate pair of potential matches. (National Center for Health Statistics, 2018a, p. 8). Researchers may request to access this variable and probabilistic weight values in order to modify the existing NHCS match criteria to adjust linkage certainty values or to conduct sensitivity analyses of assigned vital status.

To ensure confidentiality of data, NCHS provides safeguards including the removal of all personal identifiers from analytic files. Additionally, the files containing the linked 2014 NHCS-2014/2015 NDI data are only made available for research use at one of the NCHS Research Data Centers (RDCs) or one of the Federal Statistics Research Data Centers (FSRDCs) located across the country. (National Center for Health Statistics, 2018a, p. 6). A researcher who wishes to gain access must submit a proposed research project, which NCHS evaluates for feasibility and disclosure risk.

To aid researchers in evaluating whether the linked 2014 NHCS-2014/2015 NDI data will meet their analytic needs and for developing RDC proposals, data dictionaries are made available online; these documents were sent to users as part of the Transparency Reporting Project. Figure 3.2 replicates a portion of the documentation. In addition to basic fields such as Name, Label, and Description, the documentation includes other key pieces of information captured in the Notes field.

Figure 3.2 Extract from Mortality Variables Data Dictionary, 2014 NHCS-2014/15 NDI Data File

NHCS Linked to NDIMortality Variables
Date Created: 18JAN 2019
Number of Variables; 65

Variable Name	Variable (VAR) Label	VAR Type	Range of Values	Value Description	Notes
ELIGSTAT	Linkage Eligibility Status	Num	0 1	0: Ineligible 1: Eligible	Survey participants are defined as ineligible for linkage if they had insufficient identifying data to create a NHCS submission record. Please note that all survey participants are included on the linked NHCS files regardless of linkage eligibility.

Source: (National Center for Health Statistics, 2018b, p. 1)

Cost of Documentation. The estimated cost for the documentation supporting the linkage between NHCS and NDI was about $250,000, inclusive of labor hours and overhead. The labor hours included work by on-site contractors, NCHS staff, branch chief review and division clearance review. For this project, NCHS created automated codebooks in SAS which required time to format properly and to make compliant with Section 508 of the Rehabilitation Act. The project used a new linkage methodology that had to be explained for the first time in the methods and analytic guidelines that accompanied the file. The documentation text was original text that had extensive initial writing and review costs.

Responses from the Customer Survey. Overall, users indicated that they were highly confident in the linked NHCS-NDI data, with a rating of 4.8 on a 5-point scale, and that documentation effectively addressed questions. Users had positive assessments, for example, of the documentation's explanations of how sources were integrated (item #35 in Appendices 4 or 5; C3 on the researcher instrument in Appendix 3) and how data were adjusted to prevent disclosure (#55; D7), with ratings of 4.7 or higher on a 5-point scale. At the same time, free-response comments from some users pointed to ways that the documentation could potentially be improved.

In answer to the item "How did you use the data documentation or other information provided by the agency to inform or address your information needs?" (#19; A19), two users reported learning that sampling weights were not assigned to the observations. Another reported using the documentation to become familiar with how the linked data were created, which enabled the user to navigate the codebooks and understand their information.

The item "What information on data sources used in {Product} did you find useful?" (#26; B7) elicited the response that "The descriptions at the beginning of the white paper of each data source were useful for identifying the content of each dataset and why the linkage of these data sources would provide important and useful new insights." Another commented that it could be difficult to garner information on details about data collection practices.

The item "What suggestions do you have for how to improve documentation about source selections made for {Product}?" (#27; B8) generated two suggestions by respondents to the survey concerning the agency website. The methodology report could include links to the website for a researcher to learn more about the NHCS, which was described as a relatively new survey data collection and therefore potentially less familiar to researchers. Another suggestion was to expand the NHCS website to include more information about the survey data and how researchers can use them. A comment about a single variable, but an important one, concerned date of death, noting that the date of death can be collected from more than one source (linked NDI records or hospital claims records). If dates of death from the mortality and hospital record differ the data quality dimension of coherence would be affected. However, while similar information could be collected from different sources, NCHS provides users with descriptions for the methodology used to obtain date of death, indicating that the more accurate date of death is from the linked NDI record (if a linked NDI record is available) and secondarily from the hospital record if no NDI record was linked to the hospital record.

When asked for suggestions on how to improve documentation about procedures used to integrate information from different sources (#24; C11), one user asked for more information on how two years of NDI data were integrated with a single year of survey data from the NHCS. Another suggestion was to expand discussion on how national data (NDI) were integrated with data that were not nationally representative (NHCS). Another user provided the comment that this section of documentation is relatively less important.

In answer to the item on improving documentation on how to properly use the product (#46; D12), the suggestion arose that the documentation should be clearer that the data are not nationally representative.

For suggestions on improving the statistical product itself or its quality, (#65; F1), a suggestion was offered to add a flag in the survey dataset that could be used to filter duplicates of encounter-level records.

3(c). VetPop2016

The Statistical Product and its Data Sources. VetPop2016 is an actuarial projection model for the veteran population from Fiscal Year (FY) 2015 to FY2045. Using veteran data at the end of FY2015 as the base population, VetPop2016 projects the numbers of living and deceased veterans in total and by key demographic characteristics such as age, gender, period of service, and race/ethnicity. (Ahn, p. 1). The model also has the capability of making projections for the next 30 years at various geographic levels.

VetPop2016 is the eighth generation of the projection models that incorporates improvements in data, methodology, and modeling processes (Ahn, pp. 1-2). Like the previous Veteran Population Projection Model 2014 (VetPop2014), it is a geographically bottom-up model that projects future veteran populations starting at the county level. The county-level projections are then aggregated to provide veteran information at larger geographic units such as congressional districts, Veterans Integrated Service Networks (VISN), states, and at the national level.

VetPop2016 relies on information drawn from existing sources, including veteran record-level administrative data and nationally representative survey data. It does not have a unique data collection operation but instead re-uses and integrates data originally collected for other purposes. The model has four main components: the baseline data, a separation module, a mortality module, and a migration module, each of which draw upon different data sources. Table 3.2 summarizes data sources for the modules:

Table 3.2. Sources of Data used in VetPop2016, by Module

Module	Purpose	Data Sources
Separation	Project new entrants to the Veteran population	Dept. of Defense, Office of the Actuary
Mortality	Project mortality rate	Dept. of Veterans Affairs Social Security Administration
Migration	Determine county to county migration	American Community Survey, U.S. Census Bureau Internal Revenue Service commercial databases

Source: Ahn (2017, pp. 1-2), Predictive Analytics and Actuary (2017, p. 7)

The data sources for VetPop2016 include U.S. Census Bureau survey data and a database known as the United States Veterans Eligibility Trends and Statistics (USVETS) 2015. USVETS contains data acquired from over 35 sources, including the Decennial Censuses of the U.S. Census Bureau, the Department of Defence (DoD), Veterans Health Administration (VHA), National Cemetery Administration (NCA), Veterans Benefits Administration (VBA), and Social Security Administration (SSA). Some data are extracted from operational and transactional systems such as the Defence Manpower Data Center (DMDC) and the Veterans Assistance Discharge System (VADS). The VA/DoD Identity Repository (VADIR) data base was established to support the One VA/DoD data-sharing initiative to consolidate data transfers between DoD and VA. These systems directly capture real-world events and interactions with service members, veterans, and their beneficiaries.

Documentation that Communicates Data Quality to Users. The VetPop2016 users who received the Transparent Reporting Project's customer survey also received two documents—an Executive Summary (Ahn, 2017) for the model and an Overview (Predictive Analytics and Actuary, 2017) that describes data sources, model processes, and model output.

A report at NCVAS evaluates the data quality of the baseline data in USVETS. While the report was under review as of mid-2019 and not available for the public, its use within NCVAS demonstrates the recognition that assessing data quality issues is important to maintain and improve USVETS and the associated VetPop models. It is anticipated that release of the report should improve user experience with the data.

Assessing and Describing Quality. Data quality at NCVAS, as at most federal agencies, is designed to increase the reliability, accuracy, and timeliness of data to address information needs of users. Data quality is necessary to support the decision-making process and improve service outcomes. Formal rules and evaluation procedures are essential to integrate data from many sources for USVETS and VetPop2016. Evaluation of data quality is an ongoing process because it is necessary to do more than address a sudden, serious data failure. To preclude such failures, there is a need to baseline the current state of data quality and to determine improvement targets and strategies in an environment of ever-changing data.

The VA Data Governance Council has adopted seven data quality dimensions—accuracy, completeness, consistency, traceability, uniqueness, validity, and timeliness. At VA, the term "data quality dimension" means something relating to a data item, record, data set, or database that can either be measured or assessed to understand the quality of data. NCVAS used these data quality dimensions to evaluate the USVETS data quality and to determine improvement goals based on the impact of poor data quality in terms of cost, reputation, and regulatory compliance to VA. The improvement goals within these dimensions need to be balanced with the importance and intended use of the data concerned. Data may be considered 'fit for purpose' even if one or more of the above characteristics are not fully met.

A notable feature of the VA quality dimensions is their close alignment with dimensions that have been sprung from analysis of survey data. To establish the correspondence, Table 3.3 provides the definition and assessment criteria for the seven dimensions, which ranks the dimensions roughly from most important to less important.

[Table follows on next page]

Table 3.3 Data quality dimensions evaluated when constructing USVETS, VetPop and other NCVAS data

Data quality dimension
1. Accuracy - a. Definition - The degree to which a data value, or set of values, correctly represents the attributes of the real-world object or event. b. Assessment criteria - When data measures what they are intended to measure. Accurate data minimize errors (e.g., recording or interviewer bias, transcription error, sampling error) to a point of being negligible.
2. Completeness - a. Definition - The degree to which all required data are known. This includes having all required data elements, having all required records, and having all required values. b. Assessment criteria - Completeness is the extent to which all need data are available. It is usually described as a measure of the amount of available data from a statistical system compared to the amount that was expected to be obtained. At the subject area or information class, completeness refers to the specific data elements and the metadata contained within the subject area or information class assets. Within the VA, the Data Governance Analytics (DGA) unit sets acceptable levels of completeness for components of USVETS and VetPop2016 (and other data products) and documents these levels in completeness reports.
3. Consistency - a. Definition - The degree to which a set of data is equivalent in redundant or distributed databases and data elements align with documented data types and acceptable values. b. Assessment criteria - The degree to which a set of data is equivalent in redundant or distributed databases and data element values are in line with the acceptable values defined in the USVETS data dictionary.
4. Traceability - a. Definition - The extent to which data are well documented, verifiable, and easily attributed to a source. b. Assessment criteria - Clarity of the description of the source of a data element in documentation including metadata information.
5. Uniqueness - a. Definition - The degree to which no entity exists more than once within a data set. b. Assessment criteria - Data set check for case duplication.
6. Validity - a. Definition - The degree to which the data conform to defined business rules. b. Assessment criteria - Data are valid if they conform to their definition rules as to the allowable types (string, integer, floating point, etc.), formats (length, number of digits, etc.) and ranges (minimum, maximum, etc.).
7. Timeliness - a. Definition - The degree to which data are available when internal/external customers or processes require them. b. Assessment criteria - Are data released when customers require them and when forecast for release (punctuality).

The VA data quality dimensions of accuracy and timeliness are dimensions also used in the Transparent Reporting Project. In Table 3.3, consistency can be seen to be like coherence. Though not explicitly noted in the VA dimensions, relevance and comparability can also be important dimensions for VetPop2016 and USVETS.

The relevance of the VetPop2016 model is explained in the Overview, which describes the model as the source for the "latest official Veteran population projection" and how the model is used for "strategic, policy planning, and budgeting within VA and by external organizations such as other federal agencies, Congress, state governments and other organizations." (Predictive Analytics and Actuary, 2017, p. 1).

The accuracy of the model's projections is enhanced by its use of administrative data sources compiled for USVETS. Data from operational and transactional systems such as DMDC and VADS are used for auditing purposes. The system are themselves regularly audited, helping to ensure their accuracy for evaluating characteristics of the veteran population for most purposes. While no data lack errors, USVETS may be considered the gold standard for veteran data.

DoD administrative data for older veterans who served prior to the mid-1970s are incomplete due to a fire that destroyed records. These missing data have effects on accuracy. The Overview explains that to supplement VA and DoD data for older veterans, American Community Survey (ACS) data were used (Predictive Analytics and Accuracy, 2017, p. 3).

While making use of ACS data is intended to reduce errors due to missingness, the ACS data raise data quality challenges of their own, as does the use of the 2000 and 2010 Decennial Census data. The issue can be described in terms of reduced coherence: internal differences in the NCVAS data emerge due to differences in definitions and methodologies between DoD data and the Census Bureau data. In contrast to the DoD, which provides administrative data, the Census Bureau uses self-reports of veteran status on its surveys and censuses. Self-reports introduce possibilities of misreporting, which can be taken as a form of measurement error that reduces accuracy.

Steps taken in the process of streamlining data and modeling can affect the accuracy of the final projections by introducing modeling errors. For example, the VetPop2016 mortality projections were based on credibility blending of the mortality rates between VA and the Social Security Administration (SSA) and then the blended rates were smoothed (Predictive Analytics and Accuracy, 2017, p. 1). Some errors are introduced deliberately to reduce the possibility of disclosure. For example, the Census Bureau rounds highly granular data on the numbers of veterans before forwarding the file to VA in order to limit the risk of re-identifying individuals (Predictive

Analytics and Accuracy, 2017, p. 6) A form of error that is specific to linked data is a match error, by which a person identified as a veteran in one data file is mistakenly thought to be the same person in another file when the two files are matched. NCVAS matched veteran data from VA to the annual IRS tax filing data in order estimate the number of living veterans at the beginning of each FY and to estimate county-level migration patterns.

The value of the VetPop models is enhanced when they are kept up-to-date, improving their timeliness The first generation, VetPop2000, was succeeded by newer models in 2001, 2003, 2004, 2007, 2012, 2014, and 2017 (Predictive Analytics and Actuary, p. 1). The VetPop2016 model was available in June 2017.

Users have easy accessibility to the output generated by VetPop2016, which is available along with the Overview on the NCVAS website (National Center for Veterans Analysis and Statistics, 2019). As explained in the Overview the output is available as both SAS files and Microsoft Excel PivotTables (Predictive Analytics and Actuary, 2017, p. 6). The output files contain the projections at national, state, congressional district, VISN, and county levels. This form of output provides flexibility to users in the choice of how to display the data to suit their purpose.

NCVAS does not independently validate the source information. For example, NCVAS relies on DoD and the Census Bureau to perform quality assessments of their data before forwarding to NCVAS.

Costs of Documentation. A rough estimate for the cost of documentation of the VetPop2016 model is about $87,000 as measured by staff time devoted to materials that are primarily internal for VA officials and customers. There were separate additional costs for the report, currently in review, that evaluated data quality of USVETS.

Responses to the Customer Survey. Both informed consumers and researchers participated in the customer survey for the VetPop16 model. The highlights of responses below contain item codes that refer to each of the instruments in Appendices 2 and 3.

In response to the item, “How did you use the data documentation or other information provided by the agency to inform or address your information needs?” (#19 in Appendices 4 or 5; A19 on the informed consumer instrument in Appendix 2; A19 on the researcher instrument in Appendix 3), users relied on the documentation to understand the projections, including their strengths and weaknesses. Some users were themselves expected to explain the veteran projections and how they were created, including responses to media inquiries. The documentation aided that effort, as intended. For example, one response about the use of the documentation was: “Typically I use it [the documentation] to explain how Veteran projects are created and how they relate to enrollment projects.”

The item “What information on data sources used in {Product} did you find useful?” (#26; B4; B7) elicited that users found it helpful to know the name and year of data sources, a general description of the data, and how the source was used, for example, which variables were drawn from the source.

When asked, “What suggestions do you have for improving documentation about source selections made for {Product}?” (#27; B5; B8) one response was a suggestion for a research initiative that would compare veteran statistics based on three sources—Decennial Census, sample survey (ACS) and administrative records. The goal would be to have a better understanding of the extent to which differences are due to sampling, definitions of veteran status, and self-reporting. Another suggestion was to include an appendix that gives definitions for variables. Some users would benefit from more precise identification of data sources, for example, whether data from the ACS is from a publicly available source or from a special tabulation. When data are published and posted, providing citations and links is helpful. Additional information would be welcomed on reliability, or which aspects of the information might be relatively less reliable. A request was made for a data-use guide.

When asked, “What suggestions do you have for how to improve documentation about procedures used to integrate information from different sources into {Product}?” (#32; C5; C11), one user requested more information on how the data are matched. In addition, it was suggested that documentation explain how many records were unmatched from each data source, why they were unmatched, and the effects on final estimates. Some information could be difficult for any statistical agency to provide, perhaps especially on how final estimates are affected by the matching process. Even so, a fuller description of the matching process could be considered. Another suggestion was to provide greater details on underlying assumptions, for example, whether certain rates are assumed to be constant over a 30-year period. There was a request for measures of variance, although it was recognized that such measures might not be feasible for all outputs. The desirability of sensitivity analysis was raised with a suggestion that simulations and robustness checks be conducted based on different models of the rates of migration, mortality, and separation.

References for Chapter 3 (Integration of Administrative Data and Survey Data)

National Postsecondary Student Aid Study

Cominole, Melissa, John Riccobono, Peter Siegel, and Lesa Caves. 2010. 2007-08 National Postsecondary Student Aid Study (NPSAS: 08) Full-scale Methodology Report. NCES 2011-188. U.S. Department of Education. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011188

National Center for Education Statistics. No date a. "About NPSAS, " webpage update uncertain https://nces.ed.gov/surveys/npsas/about.asp

National Center for Education Statistics. No date b. "State Oversamples," webpage update uncertain https://nces.ed.gov/surveys/npsas/state_oversamples.asp

National Center for Education Statistics. No date c. "Data Information for Postsecondary Sample Surveys," webpage update uncertain https://nces.ed.gov/surveys/npsas/datainfo.asp

National Center for Education Statistics. No date d. "DataLab," webpage update uncertain https://nces.ed.gov/datalab/index.aspx

Radwin, D., Conzelmann, J.G., Nunnery, A., Lacy, T.A., Wu, J., Lew, S., Wine, J., and Siegel, P. 2018. 2015–16 National Postsecondary Student Aid Study (NPSAS:16): Student Financial Aid Estimates for 2015–16. NCES 2018-466. U.S. Department of Education. Washington, DC: National Center for Education Statistics. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018466.

Wine, Jennifer, Peter Siegel, and Rob Stollberg. 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. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018482

Linkage between the National Hospital Care Survey (NHCS) and the National Death Index (NDI)

Bercovitz, Anita, Eric Jamoom, and Denys Lau. 2018. National Hospital Care Survey Demonstration Projects: Characteristics of Inpatient and Emergency Department Encounters Among Patients with Any Listed Diagnosis of Alzheimer Disease. National Health Statistics Reports No. 121. National Center for Health Statistics, U.S. Department of Health and Human Services. December 2018. https://www.cdc.gov/nchs/data/nhsr/nhsr121-508.pdf

National Center for Health Statistics. 2017. "National Death Index," webpage updated March 7, 2017 https://www.cdc.gov/nchs/ndi/index.htm.

National Center for Health Statistics. Office of Analysis and Epidemiology. 2018a. The Linkage of the 2014 National Hospital Care Survey to the 2014/2015 National Death Index: Methodology Overview and Analytic Considerations https://www.cdc.gov/nchs/data/datalinkage/NHCS14_NDI14_15_Methodology_Analytic_Consider.pdf

National Center for Health Statistics. Office of Analysis and Epidemiology. 2018b. Codebook for the Linked 2014 NHCS and NDI File, Mortality Variables. April 2018. https://www.cdc.gov/nchs/data/datalinkage/NHCS_2014_Codebook_Mortality_Variables.pdf

National Center for Health Statistics. Office of Analysis and Epidemiology. 2018c. Codebook for the Linked 2014 NHCS and NDI File, Death Certificate and NDI Match Variables. April 2018. https://www.cdc.gov/nchs/data/datalinkage/NHCS_2014_Codebook_Death_Certificate_and_NDI_Match_Variables.pdf

National Center for Health Statistics, 2019a. "Research Data Center," webpage update March 26, 2019. https://www.cdc.gov/rdc/index.htm

National Center for Health Statistics, 2019b. "National Hospital Care Survey (NHCS)," webpage update May 2, 2019. https://www.cdc.gov/rdc/b1datatype/dt1224h.htm

The Veteran Population Projection Model 2016 (VetPop2016)

Ahn, Peter. 2017. VetPop216: Executive Summary. Department of Veterans Affairs, Office of Enterprise Integration, Data Governance and Analytics, Predictive Analytics and Actuary. June 2017. https://www.va.gov/vetdata/docs/Demographics/New_Vetpop_Model/ Vetpop16_Executive_Summary.pdf

National Center for Veterans Analysis and Statistics. No date. Veteran Population Projections 2017-2037 https://www.va.gov/vetdata/docs/Demographics/New_Vetpop_Model/Vetpop_Infographic_Final31.pdf

National Center for Veterans Analysis and Statistics. 2019. "Veteran Population," webpage updated May 10, 2019 https://www.va.gov/vetdata/Veteran_Population.asp

Predictive Analytics and Actuary. 2017. VetPop 2016: Overview. Department of Veterans Affairs, Office of Enterprise Integration, Data Governance and Analytics, Predictive Analytics and Actuary. June 2017. https://www.va.gov/vetdata/docs/Demographics/New_Vetpop_Model/Vetpop16_Overview.pdf

4. Proprietary Data

— Telecommunications services in the Consumer Price Index (CPI) —

Bureau of Labor Statistics (BLS)

— National Household Food Acquisition and Purchase Survey (FoodAPS) —

Economic Research Service (ERS)

Private firms have a long history of collecting proprietary data that describe commercial transactions between consumers and retailers; while there are always two parties to a transaction, such data are sometimes described as either "consumer data" or "retailer data" depending on who is contacted for the data. Transactions data are typically used for commercial market research purposes, providing marketing insights to retailers and manufacturers. However, these transactions data, like proprietary more generally, also confer unique benefits when re-used by statistical agencies. These data sources have become more elaborate and detailed, providing critical information that statistical agencies may use to address a wide range of topics.

This chapter examines how the Consumer Price Index (CPI) and the National Household Food Acquisition and Purchase Survey (FoodAPS) each use proprietary data that capture retail prices and other transactions information. The CPI acquired proprietary data on telecommunications services to conduct research on its data quality and assess its potential use in the CPI. FoodAPS uses proprietary data for food prices. Examining documentation about proprietary data for the CPI and FoodAPS can have lessons for other statistical agencies because other types of proprietary data can have similar issues of data quality and documentation. For example, proprietary data in general may have incomplete coverage of the population of interest, which affects the accuracyof the data, or be limited in the scope of data elements available, such as information on product characteristics, which can affect the relevance of the data. Athe same time, each source proprietary data can be expected to have some quality issues that are unique.

This chapter differs from the other chapters in this report because the discussion of data quality issues for the telecommunications relies heavily on internal BLS documentation; in contrast, other chapters, and the FoodAPS component of this chapter, are based only on publicly available documents. The BLS internal documents, which discuss private-sector data providers by name, were not designed for general release but instead were written for BLS research and decision-making purposes. By broadly describing the BLS internal assessment of data quality of proprietary data, the chapter provides an example of how other statistical agencies may approach making their own assessments.

4(a). The CPI and Telecommunications Services

The statistical product and its data sources. One of the principal responsibilities of the BLS is to measure price changes in the economy. The CPI is a measure of the cost of a "typical" market basket of goods and services that is bought by urban consumers. Changes over time in the CPI measure inflation, that is, average change in prices. The CPI provides information on a key aspect of economic conditions for private-sector decisions and for Congress, the Administration, and the Federal Reserve to develop fiscal and monetary policy.

The CPI combines economic theory with sampling and other statistical techniques and utilizes data from several voluntary surveys (Bureau of Labor Statistics, 2018a, pp. 1-15) Currently, CPI sample selection originates with the Telephone Point of Purchase Survey (TPOPS) which collects data on where consumers purchase services and goods and amounts of expenditures. The TPOPS data are subsequently used to construct a sample of retail outlets. Then a BLS economic assistant visits a selected outlet in person, or a website, to collect a price for one of the goods or services that BLS monitors each month.

Portions of the CPI use prices for telecommunications services—a multi-billion dollar industry characterized by technological changes and expansion. The BLS defines telecommunications services to include wireless telephone services, landline telephone services, Internet services (access to the Internet), and cable and satellite television services; the BLS monthly U.S. indexes combine wireless and landline into a single index for telephone services. As of December 2018, telecommunications services' relative importance in the CPI (that is, its share of a "typical" consumer's monthly expenditures) is about 4.53 percent (Bureau of Labor Statistics, 2019a). Of this, telephone services accounts for about 2.23 percent, with wireless telephone a much larger percentage than landline services (1.66 compared to 0.57 percent, respectively). Cable and satellite television account for 1.51 percent and Internet services 0.79 percent of telecommunication.

Documentation that Communicates Data Quality to Users. The mission statement of the BLS is that the agency "serves its diverse user communities by providing products and services that are accurate, objective, relevant, timely, and accessible" (Bureau of Labor Statistics, 2018b). These products may make use of alternative (non-survey) data sources, including corporate data, secondary source data (third party data), and web scraping data (Konny et al., 2019). BLS lacks direct control over alternative data sources. When considering whether an alternative data source is fit to use in the CPI, BLS staff carefully assess data quality. Staff conduct an extensive internal research process to critically evaluate whether an alternative data source is suitable on several dimensions including accuracy and reliability, timeliness, coherence and comparability, and accessibility. Cost implications are considered as well. BLS analysts and decision makers are the target users of the BLS internal documents that summarize the research. Information for the public on telecommunications in the CPI is available in Murphy and Konny (2018), Konny et al. (2019) and Bureau of Labor Statistics (2019a).

Currently, the CPI does not have many alternative data sources used to supplement or replace price collection in production. When an alternative source is used for most or all of an item, the public is informed for transparency. For example, the price information for used cars and trucks comes from an alternative data source and is documented using a fact sheet on the BLS website (Bureau of Labor Statistics, 2019b). For telecommunications services in the CPI, survey data on prices are obtained from sampled outlets or websites (Bureau of Labor Statistics, 2019a). BLS staff are conducting research using telecommunications prices obtained from alternative data sources. Once research has been completed, if alternative data sources are used to supplement or replace survey-based telecommunications prices in the CPI, BLS will add information describing the change to its telecommunication fact sheet.

Assessing and Describing Data Quality. A key tool that BLS uses to assess the quality of alternative data is known as the "Alternative Data Qualitative Assessment Scorecard." Appendix 7 provides a blank template of the scorecard and a complementary "Alternative Data Quantitative Assessment Guide" is provided in Appendix 8. The scorecard gathers information on characteristics and features the data such as granularity, scope/coverage, sampling, data usability, and opportunities for improving data quality or reducing data collection cost or burden, as well as challenges relating to the level of detail, timeliness, acquisition cost, and necessary resources and skills. Using the scorecard, BLS staff completed separate assessments of alternative data from three data providers. All three data sources were private, third-party providers and two of them involved web scraping. The material below reviews the types of questions that the scorecard poses about such data, while omitting findings BLS developed for the data under study.

The quality dimension of relevance (of the source data for BLS purposes) is embedded throughout the scorecard. One question that is a concrete example is Question 27, which asks, "Does the alternative data provide a sufficient level of detail/description for your purposes? Is it possible to define a unique item?"

Murphy and Konny (2018) state that accuracy is the most important—and the most challenging—dimension to assess when determining whether an alternative data source is fit for use at BLS. The coverage of the alternative data—the geographic areas, outlets, and time periods from which they are drawn—can affect the accuracy of the CPI. Question 12 asks, "Are data available for the current PSUs [Primary Sampling Units]? What about new PSUs coming on as a result of a geographic revision?" Question 13 asks "Are data available for the types of outlets the CPI would typically sample from?" In addition to involving accuracy, Question 13 relates to coherence. The questions show that BLS is interested in how well the alternative data would mesh with the survey data that BLS obtains by sampling prices at retail outlets—"the types of outlets the CPI would typically sample from."

Other questions have implications for accuracy too. For example, Question 10 asks, "What level is the data presented to the user? (Ex. Individual transactions or average price per unit sold? How is a unique row of data defined? How is a unique item defined?" Question 11 asks, "Data may be aggregated by geographic, item, or time dimensions, or possibly all three. For example, quarterly sales of men's apparel by geographic region. If applicable, what level(s) is the data aggregated to?" These questions about granularity—level of detail—relate to accuracy because BLS seeks to price a particular item (such as a wireless telephone service plan with 10 gigabytes of included data and unlimited calling and text messaging). A supplier of proprietary data may have a price for an item of interest to BLS that is combined with the prices of other similar items. An example is stand-alone and bundled services for wireless carriers. Combined prices can be problematic. So too can the absence of an item. Question 14 asks whether "data are available for the complete range of items included in the ELI?"

Some questions relate to accuracy and how well the data are understood. Question 16 asks, "If the dataset is a sample (rather than the universe of sales), is the sampling method used well understood?" Question 17 asks, "If the dataset is a subset or if filters/thresholds have been placed on the data, do we understand the methods of disaggregation?"

Given that the CPI is produced each month, use of alternative data in production would require that the data be timely (available shortly after the monthly time period covered by the CPI) and punctual (delivered by the date contracted between BLS and the data supplier). The scorecard shows that the BLS carefully considers issues of timing of source data availability because that in turn has implications for the timeliness and punctuality of the release of the CPI. Question 15 asks "Are the data available on a monthly basis?" Question 19 addresses "any possible issues with the timing of delivery and incorporating the data into monthly production (if that is one of the proposed uses of the data)." Question 28 asks if the alternative data would be "delivered in a reliable and timely fashion."

The dimension of accessibility refers to how easily a user can obtain and use data. Several BLS questions recognize that accessibility involves more than simple receipt of the data—the data must be manipulated, processed, and integrated into production. Question 21 asks "Does the data need to be cleaned and/or validated before it can be used? Are additional resources needed to accomplish this?" Question 30 asks "What resources and/or skills are needed to work with the data? Do you have access to these resources/skills?"

The dimension of comparability (over time) is especially important for the CPI because it is a change in the CPI over time that is used to measure price inflation. When proprietary data is introduced into the CPI instead of survey data, then the process for measuring the CPI has changed, raising the issue of comparability, which is sometimes described as "break-in-series." A method for examining what difference is made by substituting a new data source begins by asking the potential supplier of alternative data to provide historical data. These data are substituted in lieu of the survey data the BLS has already collected to calculate an experimental CPI based on (old) alternative data. If the historical CPI results and the experimental CPI results are similar, then confidence is enhanced that the alternative data source may be usable in lieu of survey data in the future. Question 15 asks the pointed question, "Can historical data be accessed if needed?"

Other questions on the scorecard ask about issues besides data quality (although these other factors can potentially affect data quality). For example, Question 24 asks, "Is respondent burden reduced through the use of the alternative data?" Question 29 asks "What are the financial costs to acquiring the data?"

BLS research has compared the CPI and an experimental index for wireless telephone services using data from one of the alternative data sources. BLS staff reached the preliminary conclusion that "this data source can replicate data collected by BLS at reduced cost, with at least the same level of accuracy" (Konny et al., 2019, p. 15). Further research is being conducted, including examination of another data source.

Telecommunication service providers are difficult to contact. For both wireless and landline (residential) services, a majority of the CPI sample has prices obtained from websites of the service-providing firms, at their request (Konny et al., 2019, pp. 15-16). This method of data collection provides more limited information than obtaining prices through contact and discussion with knowledgeable respondents at service-providing firms. At the same time, alternative data from a third-party data provider may also provide limited information, which raises several data quality issues. Internal BLS documentation reported to its users on various data quality challenges, including issues involving:

- List prices versus transactions prices. A website can provide a list price, which may exclude fees, taxes and discounts. Such factors can make the list price differ from the transactions price that the consumer pays. The CPI and research on consumer demand for telecommunications services each benefit from having transactions prices available.
- Prices paid by new versus on-going customers. Another challenge of using prices from a website is that such prices may be quoted today for a new customer or a customer who switches from an existing plan. At the same time, many of a company's customers are on-going customers. The prices paid today by on-going

customers were established in contracts made in earlier months and those prices can differ from the prices quoted today for a new customer.

- Fixed versus dynamic pricing. A further complication is that a previous contract can have built-in changes in prices and promotions rather than a fixed price over the life of the contract. If so, then an on-going customer’s contract exhibits price changes over time. As a result, even knowing the price in the month the contract was made does not capture the price the on-going customer pays today.

Ideally, the BLS would have a detailed information on prices each month for both new and on-going customers, characteristics (services) of the plans purchased by the various customers, and information on the mix or proportions of customers who pay which prices for which plans.

The BLS seeks to obtain high-quality data in an environment of rising collection costs for survey-based data. Data acquisition from an external supplier may improve quality or achieve the same level of quality at a lower cost. However, relying on an external data provider creates new challenges: potential threats to timeliness and punctuality and potential cost increases in the future. Further research will inform BLS decisions on whether and how to incorporate various alternative data sources into the process of constructing the CPI.

Costs of documentation. A rough estimate for the cost of the documentation itself was between about 125 hours or approximately $6250 and a staff cost on the order of $50 per hour.

Responses from the customer survey. As noted above, the documentation of data quality for proprietary data on telecommunications services was internal to BLS; there were no non-BLS users to contact for administering the Transparent Reporting Project’s customer survey on evaluating documentation.

4(b). FoodAPS

The statistical product and its data sources. Serious public health concerns include high rates of obesity and diet-related illnesses (Mancino et al., 2018, p. 1). Identifying which food and nutrition policies can best improve diet quality can be informed by statistical estimates about the food environment and the foods that households acquire. Information on food sources, food items acquired, prices of items, and nutritional quality of items is provided by FoodAPS—the first nationally representative survey of American households to collect unique and comprehensive data about household food purchases and acquisitions and the factors that influence food choices. FoodAPS was co-sponsored by ERS and USDA’S Food and Nutrition Service (FNS), the agency responsible for administering the Department’s food and nutrition assistance programs of which the Supplemental Nutrition Assistance Program (SNAP, formerly the Food Stamp Program) is the largest. Findings from FoodAPS include:

- In a typical week in 2012, food was acquired from a diversity of sources: large grocery stores and supermarkets (by an estimated 87 percent of U.S. households); restaurants and other eating places (85 percent); family, friends, parties, or a place of worship (37 percent); and food pantries or Meals on Wheels (1 percent) (Todd and Scharadin, 2016, p. 12);
- On average, weekly food spending by SNAP households was $52 per person, while non-SNAP households with income less than 185 percent of poverty spent $59 per person per week, and higher-income households spent $88 per person per week (Todd and Scharadin, 2016, p. 24);
- The nutritional quality of foods differs across establishments (Mancino et al., 2018, pp. 25-27);

This chapter focuses on FoodAPS and its use of proprietary data on food prices. In addition to FoodAPS, ERS uses the data for its own policy-relevant research and makes them available to USDA's Center for Nutrition and Policy Promotion (CNPP). In turn, CNPP has combined the proprietary data on food prices with other datasets to estimate the cost of a nutritious diet at various expenditure levels (Carlson et al., 2008, p. 2). One level is the Thrifty Food Plan, which is the basis for determining the dollar values of SNAP benefits. Another level is the Liberal Plan, which the U.S. Department of Defense uses to set the Basic Allowance Subsistence rate for servicemembers. Extramural researchers also use proprietary data in a program of research to better understand food purchases, food consumption, nutrition, diet-related health conditions and costs, and various economic, social, and policy factors that influence these outcomes. In recognition that proprietary food price data are used by both FoodAPS and non-FoodAPS users, ERS invited users from both groups to participate in the customer survey of the Transparent Reporting Project.

FoodAPS integrates survey data, proprietary data, and administrative records from government agencies, including State SNAP agencies. Table 4.1 provides brief descriptions of proprietary sources used in FoodAPS, which had a Geography Component designed for researchers to examine how the local food environment affects food spending patterns.

[Table follows on next page]

Table 4.1. Proprietary Data Sources in FoodAPS

Source	Description
InfoScan	InfoScan data are available from Information Resources, Inc. (IRI). The data are initially collected by food retailers at the point-of-sale (or "check-out") using electronic terminals. Infoscan data cover a large portion of retail food sales.
TDLinx	TDLinx is a database of food retailers available from Nielsen. TDLinx provides names and characteristics of food retailers across the United States. The database is designed to provide universal coverage of grocery, club, convenience, and small-format food-selling stores, although in practice not every unit in the universe may be included.
InfoUSA restaurant database	The InfoUSA restaurant database from Information Resources, Inc. (IRI) includes street addresses for eating places.
Emergency Food System Data	The Emergency Food System Data were acquired from Feeding America, which conducts a study every few years on charitable food distribution. The data include the locations of food banks, food pantries, and soup kitchens.
State Sales Tax on Soda and Snack Foods	The Robert Wood John Foundation funds the research program Bridging the Gap, which examines policies and environmental factors affecting diet, physical activity and obesity among youth.

A key advantage of these databases is their comprehensiveness. For example, TDLinx captures much of the universe of restaurants. The InfoScan data from Information Resources, Inc. (IRI) include item-level revenues and quantities for a million food products with Universal Product Codes (UPCs) sold at participating food retailers. Infoscan records can be linked with product dictionaries, data on nutrition, and manufacturer's claims regarding food healthfulness or other attributes such as an "organic" label. The product dictionaries provide descriptions that include flavor, brand, style, and type.

Documentation that communicates data quality to users. Several ERS reports explain features and quality of FoodAPS data and proprietary data. Two key examples are the FoodAPS User's Guide (Economic Research Service, 2016) and the ERS report Understanding IRI Household-Based and Store-Based Scanner Data by Muth, et al. (2016). These two reports were circulated in the Transparent Reporting Project to selected users and are the focus here.

Assessing and Describing Data Quality. When considering whether FoodAPS data are relevant for the user's application, the user may want to begin with two key types of information: What is the purpose of the data collection, that is, what types of questions or topic can be studied using the data? Second, what variables are contained in the dataset?

The User's Guide begins by stating that FoodAPS "collected information food purchased or otherwise acquired, and the prices of and nutrient characteristics of those foods, for a nationally representative sample of U.S. households. Data on factors expected to affect food acquisition decision, such as the household size and composition, demographic characteristics, income, participation in Federal food assistance programs, and dietary restrictions, were also collected" (Economic Research Service, 2016, p. 1). An ERS webpage on FoodAPS gives examples of broad research topics that can be examined using the data: "The interrelationships between American households' food acquisitions, factors influencing food demand, and household well-being," and "How access to various types of food stores is related to food choices, food security, health, and obesity" (Economic Research Service, 2019a). FoodAPS codebooks list specific variables for researchers to examine.

The relevance of the proprietary data for their use in FoodAPS and for other food-related research is explained in Muth, et al. (2016). The authors review the ERS acquisition data from IRI, a market research company, and the intended purposes of the data for food policy research (Muth et al., 2016, pp. 1-3).

FoodAPS is released as a set of public-use files (PUFs) and a set of restricted-use files (RUFs). The RUFs are available for approved researchers who sign a pledge in accordance with the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) and are subject to penalties for non-compliance. (Economic Research Service, 2016, p. 2). In addition, approved researchers agree to have ERS review results for risk of disclosure prior to public dissemination. To reduce the risk of disclosure of personal identities data in the PUFs, some variables are subject to coarsening and data swapping. (Economic Research Service, 2016, pp. 15-19). As a result, statistical estimates from the PUFs and RUFs can differ slightly. Making PUFs available increases accessibility of the data while protecting confidentiality and public trust. At the same time, statistical techniques that reduce disclosure risk can have implications on the accuracy of the data and statistical results. For example, data swapping can affect the weighted distributions and multivariate relationships, as noted in the User's Guide which also points out that the challenge of promoting use while protecting confidentiality is faced by all surveys.

Accuracy can be considered in terms of sampling and nonsampling errors. Estimated variances are a measure of sampling error. As a survey based on a complex sample design, FoodAPS provides weights and variables for estimating variances using either Taylor series or jackknife repeated replication (Economic Research Service, 2016, p. 13). The User's Guide provides an appendix with code (in SAS, Stata, and R) for implementing variance estimation. Non-sampling errors may be more challenging to measure than sampling error. The User's Guide considers three sources of non-sampling errors—survey non-response, underreporting of food acquisitions, and observational effects by which households change their food purchase behaviors due to participating in the survey, and (Economic Research Service, 2016, pp. 24-29). The User's Guide reports two types of response rates for FoodAPS (Economic Research Service, 2016, p. 11) and the results of a non-response bias analysis. Item nonresponse in FoodAPS is reported to be "generally low" (Economic Research Service, 2016, p. 24); FoodAPS provides information on item nonresponse for every variable in data file.

The proprietary data used in FoodAPS also have data issues concerning accuracy. Perhaps most importantly, the data in InfoScan are collected from stores that choose to have agreements with IRI. The types of stores include grocery, drug, convenience, mass merchandiser, club, dollar, and defense commissary stores (Muth et al., 2016 p. 19). In 2012, over 41,000 stores were covered, as were billions of transactions. However, sales from smaller, independent stores may not be well represented. Muth et al., (2016, p. 41) note that this feature of the data can make the data problematic for some types of analyses, giving WIC purchases as an example. Muth et al. (2016, page) also note that data for private-label (store-brand) products are not as complete as for branded products, reflecting that food retailers must approve whatever information is released by IRI. Muth et al. (2016, p. 42) report that InfoScan data are not weighted, which precludes calculation of a nationally representative price; the data are representative of those stores that self-select to be included.

Similarly, a limitation of the TDLinx store database is its defined universe of stores. The TDLinx database, by design, does not cover grocery stores having under $1 million in sales. This omits a distinct segment of the food retail landscape and, therefore, the dataset does not include the total universe of food retailers.

Like other micro-level datasets released by federal statistics agencies, FoodAPS contains imputed data for such items as income. Besides the accuracy of imputed data, the accuracy of self-reported biological data for height, weight and the resulting BMI can be questioned; FoodAPS contained flags for biologically implausible values (Economic Research Service, 2016, p. 39).

Timeliness and punctuality have been considered especially important to users inasmuch as data cannot be used until they are released (Biemer and Lyberg, 2003, p. 17) The timeliness of the FoodAPS data may be measured as the lag between the time when data were collected and the time when data were available for research. FoodAPS was fielded from April 2012 to mid-January 2013. RUFs became available in 2014, while PUFs were available later, in 2016, due to the time necessary to implement disclosure protections.

To promote accessibility and clarity (interpretability), ERS has a portfolio of reports that document the collection, use and quality of FoodAPS data. The ERS webpage on FoodAPS Data Quality and Accuracy contains links to six data quality reports on: lessons learned from designing and conducting FoodAPS; instrument design, response burden, use of incentives, and response rates; sample design; completeness and accuracy; potential for nonresponse bias; imputation approaches for income and price data; and a review of FoodAPS from a data user's perspective (Economic Research Service, 2019c). While FoodAPS was designed and conducted by Mathematic Policy Research under contract to ERS, ERS contracted with Westat, Inc. to conduct an independent review and prepare most of the reports listed above. The report by Wilde and Ismail (2018) that considered a data user's perspective is notable in that, like the Transparent Reporting Project, it conducted a survey of users—in this case, a survey of research teams that had used FoodAPS. An ERS webpage on Documentation contains links to interviewing documentation, respondent confidentiality forms, household screening tools, food books, and the User's Guide and a dozen codebooks (Economic Research Service, 2019b).

Because FoodAPS has been administered once, the issue of comparability of data across survey administrations does not arise. However, it is worth noting that just prior to conducting FoodAPS, ERS transitioned from using retail food price data provided by Nielsen to data provided by IRI. This switch between the two data suppliers was prompted by re-competing the contract by which retail food price data are secured. Statistical agencies that use proprietary data face a potential break-in-series when the agency shifts between data suppliers or when a data supplier discontinues a series.

FoodAPS used a variation on the definition of a "household" that the U.S. Census Bureau (Census) and other agencies use for household surveys. A "Census household" is an address-based concept: those people who reside at a given address, regardless of whether there are family connections among them all (Census Bureau, undated) In contrast, a "FoodAPS household" is "all persons who live together and share food and who expect to be present at the sampled address during at least part of the data collection week." (Economic Research Service, 2016, pp. 2-3) This definition was adopted to reflect the definition of a SNAP unit ("SNAP household") and to "match food acquisitions as closely as possible with the people at the sampled residence during the week." Because this definition does not match the Census definition, coherence between FoodAPS results and results of Census surveys is impinged.

A general drawback of obtaining data from commercial vendors concerns the completeness of the documentation because of the proprietary nature of the data. Documentation can be limited relative to the documentation available for public data from federal statistical agencies. Two other limitations can have effects on relevance, due to limited granularity. First, some retailers release data for aggregated market areas (rather than for each individual store) for data privacy. Because these geography-based aggregations vary by retailer, it can be difficult to examine geographic variation or conduct analyses by State or other detailed geographic areas for certain retailers (Muth et al., 2016, p. 19) Second, some retailers do not release private label (store brand) data or release it at an aggregate product level.

As part of its research on data quality, ERS has examined comparability of statistical results obtained from FoodAPS with results from the Consumer Expenditure Survey (CE), which BLS conducts to collect information on expenditures across consumer products including foods, and from the National Health and Nutrition Examination Survey, which is conducted by the National Center for Health Statistics (Clay, et al., 2016). To understand the coverage, representativeness, and limitations of proprietary data on food retail prices, ERS commissioned a set of studies on the characteristics and statistical properties of the scanner data available from Nielsen and IRI. Studies by Einav, et al. (2008a), Einav, et al. (2008b), Einav, et al. (2010) and Muth, et al. (2016) examined the coverage and accuracy of scanner data. Zhen, et al. (2009) and Sweitzer, et al. (2017) compared household scanner data, obtained from a large panel of participating households, with data from the CE. These studies found that scanner data and associated files can be an extensive resource for consumer food and health research, but researchers must understand the complexity, properties, and limitations of the datasets. Statistical agencies and data users considering the quality of proprietary data for prices of non-food products may wish to review some of the food price literature and either locate or develop comparable analyses of proprietary data on non-food items.

Costs of documentation. The report by Muth, et al. (2016) is estimated to have cost about $140,000 based on contributions to the report by both extramural and ERS researchers. An ERS extramural agreement supported a portfolio of external work that included the report. A rough ex post allocation of the agreement's total cost results in an estimate of about $65,000 as the report's extramural cost component. In addition, about $75,000 of staff time internal to ERS was devoted to the report, for a total of $140,000. The cost of the User's Guide was estimated to be about $30,000 of ERS staff time.

Responses from the Customer Survey. For proprietary data from IRI, the Transparent Reporting Project invited two subgroups of users: FoodAPS users (FoodAPS included IRI data) and those who used IRI data without FoodAPS. Respondents to the survey were researchers from both subgroups, which are identified as "FoodAPS" and "IRI Data" in Appendix 5. The FoodAPS users received two documents to evaluate—the User's Guide and Muth, et al (2016), while the IRI Data users received only Muth, et al. (2016) because the User's Guide for FoodAPS was not relevant for them. Several highlights of their responses are noted below.

Like other respondents to the customer survey, the FoodAPS and IRI users considered the data quality to be relatively strong for relevance and accuracy, less strong for timeliness and access. FoodAPS users had relatively high ratings for quality of documentation.

The users' responses to the free-response items, which are reviewed below, were quite informative. In one sense, the user responses are specific to FoodAPS and IRI data. However, in another sense the responses are more universal, reflecting the kinds of issues that users have when accessing many products and understanding documentation when proprietary data are involved.

When asked, “How did you use the data documentation or other information provided by the agency to inform or address your information needs?” (#19 in Appendices 4 or 5; A19 on the informed consumer instrument in Appendix 2; A19 on the researcher instrument in Appendix 3), the responses indicate that the documentation was serving the purposes that were intended. Users referred to the documentation to determine: whether to use, for their purposes, the scanner data from retailers or from the panel of consumers; variable names; how best to analyze the data; and how to explain the data to others. One respondent noted that the documentation was helpful for writing grant proposals. While another respondent wrote that the documentation is “well-crafted and fairly easy to follow,” the user also noted that it was challenging to identify which documentation file had which information. A user noted that a difficulty of using restricted-use data is that some of its documentation is not available outside of the secure data environment.

For the item “What information on data sources used in {Product} did you find useful?” (#26; B7), responses referred to how FoodAPS was linked to SNAP administrative records and that documentation on those linkages was useful. So too was documentation on store locations and food prices.

A key purpose of the Transparent Reporting Project was to identify ways to improve documentation. A benefit of the customer survey is that it posed that very question in general terms to the users: “What suggestions do you have for how to improve documentation about source selections made for {Product}?” (#27; B8). One user noted that it was challenging to understanding the linkages between FoodAPS and SNAP administrative records, which is the same topic that was reported above that was found to be “useful.” These user responses may not be contradictory: users can benefit from information on a topic and still seek additional detail because they find the information useful—not despite it. A user requested fuller documentation on the construction of food price data; at the same time, the user did deduce the needed information from coding, which points to how it can be useful for a statistical agency to make such computer code public when possible. Another user suggested that food products be flagged if they are under USDA labeling jurisdiction which, together with other suggestions, would support research on determining costs and benefits for new labeling regulations.

Another free-response item asked, “What suggestions do you have for how to improve documentation about procedures used to integrate information from different sources into {Product}?” (#32; C11) A user responded that while “the FoodAPS team did great work,” it was difficult to match stores between FoodAPS and proprietary data sources.

Another documentation question was: “What suggestions do you have for how to improve documentation about how to properly use the product?” (#46; D12) Users suggested that the agency: provide code to impute missing information and to merge datasets; add an appendix explaining variables; explain upfront specific changes to data that were done to ensure confidentiality; and adding more data, e.g., vehicular ownership by households and certain characteristics about food retailers.

ERS users were asked a question designed specifically for them about possible gaps in documentation: "What issues or questions related to the quality or statistical properties of the IRI Data or the appropriateness of using them for specific areas of research are not currently addressed in the enclosed documentation that would be useful to potential IRI users?" (#67; G2) Users sought more information on the meaning of some variable names, information on imputation, and differences in certain product codes.

There was another key question that asked about improving data quality rather than documentation: "In contrast to improving documentation, do you have suggestions for improving the statistical product itself or its quality?" (#65; F1) Concern was expressed about the quality of the data on SNAP participation. It was urged that FoodAPS verify participation in WIC and other food programs. Another suggestion was to make available practice datasets for students and sample code from published studies. One user offered suggestions for obtaining more information about veterans' status, disabilities, and employment, and then wrote, "Overall though, this is an amazing resource!"

References for Chapter 4 (Proprietary Data)

Telecommunications in the Consumer Price Index

Bureau of Labor Statistics. 2018a. "Chapter 17. The Consumer Price Index, (updated 2-14-2018)" in *Handbook of Methods* https://www.bls.gov/opub/hom/pdf/cpi-20180214.pdf

Bureau of Labor Statistics, 2018b. "Mission Statement," webpage updated June 7, 2018. https://www.bls.gov/bls/blsmissn.htm

Bureau of Labor Statistics, 2019a. "Measuring Price Change in the CPI: Telecommunications Services." https://www.bls.gov/cpi/factsheets/telecommunications.htm

Bureau of Labor Statistics, 2019b. "Measuring Price Change in the CPI: Used Cars and Trucks." https://www.bls.gov/cpi/factsheets/used-cars-and-trucks.htm

Konny, Crystal, Brendan Williams, and David Friedman. 2019. "Big Data in the U.S. Consumer Price Index: Experience & Plans," presented at the Conference on Research in Income and Wealth (CRIW) Conference *Big Data for 21st Century Economic Statistics*, March, 2019. https://papers.nber.org/sched/CRIWs19

Murphy, B. and C. Konny. 2018. "Quality Considerations for Administrative Data Used for the Producer Price Index (PPI) and Consumer Price Index (CPI) development," in Brown A., A. Caporaso, K.G. Abraham, and F. Kreuter, *Findings from the Integrated Data Workshops hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society* https://nces.ed.gov/fcsm/pdf/Workshop_Summary.pdf

Food Prices Data in the National Households Food Acquisition and Purchase Survey

Biemer and Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons.

Carlson, Andrew, Mark Link, WenYen Juan, Kristin Marcoe, Lisa Bente, Hazel Hiza, Patrician Guenther, and Ephraim Leibtag. 2008. *Development of the CNPP Prices Database*. CNPP-22. USDA, Center for Nutrition Policy and Promotion. May 2008. http://ageconsearch.umn.edu/record/45851/files/PricesDatabaseReport.pdf

Clay, Dawn Marie, Michele Ver Ploeg, Alisha Coleman-Jensen, Howard Elitzak, Christian Gregory, David Levin, Constance Newman, and Matthew Rabbitt. 2016. *Comparing National Food Acquisition and Purchase Survey (FoodAPS) Data With Other National Food Surveys' Data*. Economic Information Bulletin No. 157. USDA, Economic Research Service. July 2016. https://www.ers.usda.gov/publications/pub-details/?pubid=79892

Economic Research Service, U.S. Department of Agriculture. 2016. *National Household Food Acquisition and Purchase Survey (FoodAPS): User's Guide to Survey Design, Data Collection, and Overview of Datasets*. November, 2016. https://www.ers.usda.gov/media/8804/0_foodaps-user-guide-puf.pdf

Economic Research Service, U.S. Department of Agriculture. 2019a. "Background," webpage updated April 2, 2019.
https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/background/

Economic Research Service, U.S. Department of Agriculture. 2019b. "Documentation," webpage updated April 2, 2019.
https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/documentation/

Economic Research Service, U.S. Department of Agriculture. 2019c. "Data Quality and Accuracy," webpage updated April 4, 2019.
https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/data-quality-and-accuracy/

Einav, Liran, Ephraim Leibtag, and Aviv Nevo. 2008a. *Not-so-Classical Measurement Errors: A Validation Study of Homescan.* NBER Working Paper No. 14436. October 2008.
https://www.nber.org/papers/w14436

Einav, Liran, Ephraim Leibtag, and Aviv Nevo. 2008b. *On the Accuracy of Nielsen Homescan Data.* Economic Research Report No. 69. USDA, Economic Research Service. December 2008.
https://www.ers.usda.gov/publications/pub-details/?pubid=46114

Einav, Liran, Ephraim Leibtag, and Aviv Nevo. 2010. "Recording discrepancies in Nielsen Homescan data: Are they present and do they matter?" *Quantitative Marketing and Economics*, 8(2), 207-239.
https://ideas.repec.org/a/kap/qmktec/v8y2010i2p207-239.html

Mancino, Lisa, Joanne Guthrie, Michele Ver Ploeg, and Biing-Hwan Lin. 2018. *Nutritional Quality of Foods Acquired by Americans; Findings from USDA's National Household Food Acquisition and Purchase Survey* Economic Information Bulletin No. 188. USDA, Economic Research Service. February 2018.
https://www.ers.usda.gov/publications/pub-details/?pubid=87530

Muth, Mary, Megan Sweitzer, Derick Brown, Kristen Capogrossi, Shawn Karns, David Levin, Abigail Okrent, Peter Siegel, and Chen Zhen. 2016. *Understanding IRI Household-Based and Store-Based Scanner Data.* Technical Bulletin No. 1942. USDA, Economic Research Service.
https://www.ers.usda.gov/publications/pub-details/?pubid=47636

Sweitzer, Megan, Derick Brown, Shawn Karns, Mary Muth, Peter Siegel, and Chen Zhen. 2017. *Food-at-Home Expenditures: Comparing Commercial Household Scanner Data from IRI and Government Survey Data.* Technical Bulletin 1946. USDA, Economic Research Service. September 2017.
https://www.ers.usda.gov/publications/pub-details/?pubid=85251

Todd, Jessica and Benjamin Scharadin. 2016. *Where Households Get Food in a Typical Week: Findings from UDSA's FoodAPS.* Economic Information Bulletin No. 156. USDA, Economic Research Service. July 2016.
https://www.ers.usda.gov/publications/pub-details/?pubid=80541

U.S. Census Bureau. undated. "Households and Persons Per Household," webpage about the American Community Survey.
https://www.census.gov/quickfacts/fact/note/US/HSD410217

Wilde, Parke and Mehreen Ismail. 2018. Review of the National household Food Acquisition and Purchase Survey (FoodAPS) from a data user's perspective. Prepared for USDA, Economic Research Service. February, 2018. https://www.ers.usda.gov/media/9776/foodaps_datauserperspective.pdf

Zhen, Chen, Justin Taylor, Mary Muth, and Ephraim Leibtag. 2009. "Understanding Differences in Self-Reported Expenditures between Household Scanner Data and Diary Survey Data: A Comparison of Homescan and Consumer Expenditure Survey," Review of Agricultural Economics 31 (3): 470-92, October 2009. https://academic.oup.com/aepp/article/31/3/470/8063

5. Integration of Data from Multiple Surveys

— Scientists and Engineers Statistical Data System (SESTAT) —

National Center for Science and Engineering Statistics (NCSES)

Previous chapters analyzed how agencies have communicated data quality about statistical products that are constructed from administrative and commercial sources, often in combination with survey data. The case studies examined the Personal Consumption Expenditures component of Gross Domestic Product, combinations of survey and administrative data (chapter 3), and the use of proprietary data (chapter 4). The final case study in this chapter examines a statistical product constructed from multiple sample surveys—the Scientists and Engineers Statistical Data System (SESTAT).

SESTAT is developed and disseminated through the National Center for Science and Engineering Statistics (NCSES) of the National Science Foundation (NSF). An expert panel convened by the Committee on National Statistics of the National Academies of Sciences, Engineering, and Medicine considered the NCSES surveys to be an international "gold standard" on the science and engineering workforce "as the result of an active program of evaluation and improvement" (National Academies of Sciences, Engineering, and Medicine, 2018, p. 1).

Recent findings from SESTAT include:

- "From 2003 to 2013, the number of scientists and engineers residing in the United States grew from 21.6 million to 29.0 million" (Lan et al., 2015, p. 1).
- "An important factor in this growth has been immigration: In 2013, 18% (5.2 million) of the scientists and engineers residing in the United States were immigrants, whereas in 2003, 16% (3.4 million) were immigrants. In 2013, the majority of U.S. immigrant scientists and engineers were naturalized U.S. citizens (63%), whereas 22% were permanent residents and 15% were temporary visa holders (Lan et al., 2015, p. 1).

The statistical product and its data sources. Beginning in the 1950s, the federal government prioritized production and dissemination of information about the science and engineering workforce in the United States largely for military preparedness purposes (National Academies of Sciences, Engineering, and Medicine, 2018, pp. 9-10). Responsibility for collecting and disseminating information on this population was embedded in the NSF upon its formation in 1950 (National Research Council, 1989, pp. 21, 59, 62). Initially, NSF met this mandate in two ways. During the 1950s, NSF assumed collection of the National Register of Scientific and Technical Personnel which had been started by the Federal Security Agency in 1952. The agency also provided the National Research Council with funding to support the Survey of Earned Doctorates (SED), which had started in 1946 to obtain data the number of new doctoral-level research scientists entering the workforce on a regular basis. Interest in the science and engineering population has remained high as scientists and engineers have a disproportionate effect on the U.S. economy and its rate of technological change (National Research Council, 2003, p. 5).

Since the 1950s, NSF has periodically reviewed and revamped how it collects and disseminates information on the science and engineering workforce. NSF has a continuing mandate most recently expressed in the America COMPETES Reauthorization Act of 2010, which made amendments to the National Science Foundation Act of 1950 (U.S. Government Printing Office, 2010). NSF is meeting its responsibility in this research area through the SESTAT system developed by NCSES, which describes SESTAT as "a comprehensive and integrated system of information about the employment, educational and demographic characteristics of scientists and engineers in the United States" (National Center for Science and Engineering Statistics. 2017a).

The SESTAT system was first constructed in 1993, replacing the Scientific and Technical Personnel Data System, with data originally drawn from the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctoral Recipients (SDR) (National Research Council, pp. 6-7; National Academies of Sciences, Engineering, and Medicine, 2018, pp. 37-38). Long-form data from the Decennial Census formed the sampling frame of the NSCG, which interviews college graduates about work activities, salary, and demographic characteristics. Respondents were reinterviewed on a 2-year cycle during the 1990s. A new NSCG was initiated in 2003 with the 2000 Decennial Census longform data. Data from NSCG were augmented with information about new college graduates in science, engineering, and health fields who earned their degrees after the Decennial Census through the biennial NSRCG. Respondents for NSRCG were sampled from rosters for college graduates in a given year provided by sampled postsecondary institutions. Finally, to make the data representative of Ph.D. recipients, data were drawn from the biennial SDR collections. Data in SDR are representative of those holding research Ph.D.s in science, engineering, and health fields with samples drawn from the Survey of Earned Doctorates (SED). The SED is a census of all new Ph.D. earners each year.

With the advent of the American Community Survey (ACS), the Decennial longform data that had formed the sampling frame for NSCG was discontinued (National Academies of Sciences, Engineering, and Medicine, 2018, pp. 67-68). NCSES began using ACS data to construct the NSCG sampling frames. Because new ACS data are available annually, the NSCG sampling approach was modified to rotate in new sample members from ACS data every other year. This allowed for recent college graduates not available for the initial data collection period of NSCG to be integrated directly into the NSCG data collections. As a result, NSRCG data were not needed for information about college graduates who earned their degrees after NSCG sample draws and the NSRCG was discontinued in 2010.

SESTAT currently integrates data from the NSCG and the SDR, as summarized in table 5.1.

Table 5.1. Survey Data Sources in SESTAT

Source	Description
National Survey of College Graduates (NSCG)	The NSCG is a biennial panel survey of college graduates that focuses on the science and engineering workforce, collecting information on degree field, occupation, work activities, salary, and demographic information.
Survey of Doctorate Recipients (SDR)	The SDR is a biennial panel survey of research Ph.D. earners in the science, engineering, and health fields that collects data on respondent demographics, educational history, work status and occupation.

Source: National Center for Science and Engineering Statistics (2019b, 2019c)

Two widely used statistical products that are produced biennially using SESTAT data (or a component) are *Science and Engineering Indicators* and *Women, Minorities, and Persons with Disabilities in Science and Engineering*, both of which are mandated by the U.S. Congress (National Academies of Sciences, Engineering, and Medicine, 2018, p. 40).

Documentation that communicates data quality to users. Potential and current users of SESTAT are provided with information about the statistical product in several different sources. The Transparent Reporting Project provided 9 sources for users to consider when evaluating the documentation. Of these, 4 sources were publications: a Committee on National Statistics review of SESTAT in *Improving the Design of Scientists and Engineers Statistical Data Systems* (National Research Council, 2003); technical notes for the product provided in *Scientists and Engineers Statistical Data System (SESTAT), 2013, Technical Notes* (Lan, 2015); and sections on the data sources and their limitations in two NCSES publications, *Immigrants' Growing Presence in the U.S. Science and Engineering Workforce: Education and Employment Characteristics in 2013* (Lan, et al., 2015), and *Prevalence of Certifications and Licenses among the College-Educated Population in the United States* (Finamore and Foley, 2017). In addition, links for 5 webpages were provided to users as part of the customer survey: two webpages that describe SESTAT's 2010 and 2013 data (National Center for Science and Engineering Statistics, 2015a, 2015b); webpages that describe the NSCG provides its past questionnaires (National Center for Science and Engineering Statistics, 2019a, 2019b); and a webpage on Frequently Asked Questions (National Center for Science and Engineering Statistics, 2017a).

Assessing and Describing Data Quality. Publicly available information on SESTAT is spread across several documents, reflecting how SESTAT is an integrated data product. When considered in total, the information from these different sources provide considerable detail about SESTAT and the quality of its data.

The relevance of SESTAT and its component parts, the NSCG and the SDR, is made clear in NCSES publications, as well as a series of independent reports from panels convened by the Committee on National Statistics. For example, technical notes for SESTAT 2013 explain that SESTAT "provides a comprehensive picture of the number and characteristics of individuals in the United States with a bachelor's or higher-level degree, with a focus on those having science and engineering (S&E) degrees or working in S&E occupations" (Lan, 2015, p. 1). The 2003 panel report explains why that the population is important to understand and study because the work of scientists and engineers affects technological and economic progress (National Research Council, 2003, p. 5). Similar information is provided about the NSCG and SDR in study-specific webpages that cross-link with the main SESTAT web page.

Much of the information about the accuracy of SESTAT statistics is provided on the webpages dedicated to NSCG (National Center for Science and Engineering Statistics. 2019a) and SDR (National Center for Science and Engineering Statistics, 2019c) and in publications accessible through those webpages. This way of organizing documentation on SESTAT is a product of how SESTAT itself is produced. Data from SESTAT's underlying data sources provide unique and non-overlapping information about populations covered by NSCG and SDR. As a result, the accuracy of the component data sets is central to the accuracy of the SESTAT estimates. Documentation for the NSCG and SDR communicates how these statistical products, in turn, are founded on their own well understood and well evaluated sources. The SDR sample is drawn from a universe of newly awarded Ph.D. recipients who are included in the NCSES Survey of Earned Doctorates (SED) and the NSCG starts with the American Community Survey (ACS) as its base sampling frame. Information from the SED and ACS is brought forward into SDR and NSCG to reduce respondent burden.

The method of integrating into SESTAT data from NSCG and SDR builds a high degree of coherence into the integrated data product because possible inconsistencies on common constructs are effectively ruled out. The integration uses a unique linkage rule making each data source the sole source of information on specific SESTAT subpopulations. By design, "[R]espondents with doctorates in science, engineering, or health (SEH) fields from U.S. academic institutions who are identified through SED are SDR sample cases" (Lan, 2015, p. 2). For this type of researcher, SDR is the only data source incorporated into SESTAT. NSCG sample cases cover "those with research doctorates in other fields [but who work in science and engineering occupations] or those with research doctorates awarded by foreign institutions." Data about such researchers in SESTAT are drawn solely from NSCG.

Following a recommendation from the Committee on National Statistics (CNSTAT), a redesign of the SESTAT was initiated in 2013 by shifting to ACS to replace the NSRCG component. NCSES reports that the agency "evaluated the redesign of SESTAT in regard to improving timeliness, quality, efficiency, and reducing overall survey costs" (National Center for Science and Engineering Statistics, 2014). At the same time, though, the comparability over time of SESTAT estimates may be affected when a new methodology is introduced. Details on current procedures for data collection and processing are in the most recent CNSTAT review of SESTAT (National Academies of Sciences, Engineering, and Mathematics, 2018, pp. 91-108).

A priority for NCSES is promoting access to the SESTAT data, which complements the completion of its own SESTAT-based research reports. NCSES provides multiple public-facing data access tools tailored to the needs of its different users. Some users may prefer prepopulated tables while others prefer to generate their own data tables. Both options are available on the NCSES website (National Center for Science and Engineering Statistics, 2017b) In addition, public-use micro-level data can be downloaded directly from the NCSES website, which also provides instructions for how to access restricted-use data.

An interesting feature of the SESTAT documentation makes SESTAT information more accessible the SESTAT Metadata Explorer, which enables users to browse for variables within a specific user-chosen survey or across the seat of NCSES surveys (National Center for Science and Engineering Statistics, undated). This feature makes the information more transparent for SESTAT and other related NCSES data products.

Costs of documentation. The cost associated with the report on SESTAT provided by the National Research Council (2003) is $70,000. The SESTAT's Technical Notes for 2013 (Lan, 2015) were estimated to have cost $15,000. The costs of the NCSES publications by Finamore and Foley (2017) and by Lan, et al. (2015) are each estimated at $12,000, although only some fraction of that cost constitutes the specific cost of producing the documentation sections that explain the data sources and limitations.

Responses from the Customer Survey. As with the other products considered for this report, the respondents to the customer survey about SESTAT found the data product to be relevant to their research needs, reliable, consistent in terms of aligning with other sources of information on the topic and were confident in the information they received from SESTAT. One respondent made a fundamental point that applies broadly to integrated data products: "SESTAT is great because it has a larger sample size than SDR and NSCG. For issues related to college graduates and their career trajectories, SESTAT is better than [either] NSCG or SDR [alone]."

Respondents also thought the product was relatively easy to access (respondents provided average scores of 4.5 or higher on these dimensions on a 5-point scale). As with other products in the report, respondents wanted the data more quickly and more predictably (average rates were between 3.0 and 3.3 on these dimensions). Some were not aware of plans to continue the product asking, “Can you continue to pool SDR and NSCG into SESTAT? Right now, the most recent data are for 2013.”

In terms of the documentation, respondents thought what was provided was relevant, well written, and clear with scores at or near 5 for quality of the writing. However, they also indicated that more detail about how the SESTAT product is developed and evaluated would be helpful. Overall ratings for the amount of detail provided about the product in the written documentation was 3.3 This may be in part because some users of the product found it difficult to navigate to more technical information about SESTAT on the NCSES web site. One respondent indicated that, “The SESTAT page is hard to find from the NCSES home page, unless you know to look for it. More visibility of that page might be helpful.”

The concerns with the level of detail available in the written documentation may be offset by details available for those using SESTAT through personal communications with the NCSES SESTAT team. Across items about information provided through direct agency, respondents provided ratings ranging from 4.3 to 5. Satisfaction with information provided through agency contacts was also reflected in an average rating of 4.7 given to a unique NCSES-specific survey item on satisfaction with the usefulness of the secure data facility available for SESTAT analyses.

In answer to the free-response question “How did you use the data documentation or other information provided by the agency to inform or address your information needs?” (#19 in Appendices 4 or 5; A19 on the informed consumer instrument in Appendix 2; A19 on the researcher instrument in Appendix 3) users responded that documentation helped them use data correctly and report data limitations. The questionnaire and variable list were used to identify if the data included variables of interest. The question “What information on data sources used in {Product} did you find useful?” (#26; B4; B7) elicited the responses that information on SESTAT components surveys was useful, as was the codebook. To improve documentation about source selections (#27; B5; B8), one user suggested a technical report that reviewed the redesigns of the SESTAT surveys.

In response to the item “What suggestions do you have for how to improve documentation about procedures used to integrate information from different sources into {Product}?” (#32; C5; C11), one user expressed a desire for state-level data. Suggestions on “how to improve documentation about how to properly use the product” (#46; D4; D12) included more information on standard errors of the estimates from SESTAT. Another felt that the usefulness of the data was diminished by steps taken to preserve confidentiality.

One suggestion for improving the statistical product itself or its data quality (#65; F1) urged NCSES to work closely with researchers when surveys are redesigned or changed. Another suggestion was for naming conventions to be the same over time; while most variables retain the same name in different years, others have slightly different names.

References for Chapter 5 (Integration of Data from Multiple Surveys)

Finamore, John and Daniel Foley. 2017. Prevalence of Certifications and Licenses among the College-Educated Population in the United States. InfoBrief, NSF 17-312. National Center for Science and Engineering Statistics. January 2017. https://www.nsf.gov/statistics/2017/nsf17312/nsf17312.pdf

Lan, Flora. 2015. Scientists and Engineers Statistical Data System (SESTAT), 2013, Technical Notes. National Center for Science and Engineering Statistics. https://ncsesdata.nsf.gov/us-workforce/2013/ses_2013_tech_notes.pdf

Lan, Flora, Katherine Hale, and Emilda Rivers. 2015. Immigrants' Growing Presence in the U.S. Science and Engineering Workforce: Education and Employment Characteristics in 2013. InfoBrief, NSF 15-328. National Center for Science and Engineering Statistics. September 2015. https://www.nsf.gov/statistics/2015/nsf15328/nsf15328.pdf

National Academies of Sciences, Engineering, and Medicine. 2018. Measuring the 21st Century Science and Engineering Workforce Population: Evolving Needs. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/24968/measuring-the-21st-century-science-and-engineering-workforce-population-evolving

National Center for Science and Engineering Statistics. undated. "SESTAT Metadata Explorer." https://ncsesdata.nsf.gov/metadataexplorer/metadataexplorer.html?tab=howto

National Center for Science and Engineering Statistics. 2014. "Announcement [on NCSES redesigns SESTAT]," webpage updated December 16, 2014. https://www.nsf.gov/statistics/sestat/redesign.cfm

National Center for Science and Engineering Statistics. 2015a. "Scientists and Engineers Statistical Data System (SESTAT), 2013," webpage updated February 18, 2015. https://www.nsf.gov/statistics/srvygrads/#tabs-4&tabs-2

National Center for Science and Engineering Statistics. 2015b. "Scientists and Engineers Statistical Data System (SESTAT), 2015," webpage updated September 30, 2015. https://www.nsf.gov/statistics/srvygrads/#tabs-4&tabs-2

National Center for Science and Engineering Statistics. 2017a. "SESTAT—Frequently Asked Questions," webpage updated September 20, 2017. https://www.nsf.gov/statistics/sestat/#sestat-faq

National Center for Science and Engineering Statistics, 2017b. "SESTAT—Data," webpage updated September 20, 2017. https://www.nsf.gov/statistics/sestat/index.cfm#sestat-data&datatables

National Center for Science and Engineering Statistics. 2019a. "National Survey of College Graduates –Survey Description," webpage updated February 2019. https://www.nsf.gov/statistics/srvygrads/#tabs-4&tabs-2

National Center for Science and Engineering Statistics. 2019b. “National Survey of College Graduates—Questionnaires,” webpage updated February 2019. https://www.nsf.gov/statistics/srvygrads/#tabs-4&tabs-2

National Center for Science and Engineering Statistics. 2019c. “Survey of Doctorate Recipients,” webpage updated February 2019. https://www.nsf.gov/statistics/srvydoctoratework/#qs&sd.

National Research Council. 1989. Surveying the Nation's Scientists and Engineers: A Data System for the 1990s Washington, DC: The National Academies Press. https://www.nap.edu/catalog/19026/surveying-the-nations-scientists-and-engineers-a-data-system-for

National Research Council. 2003. Improving the Design of the Scientists and Engineers Statistical Data System (SESTAT).Washington, DC: The National Academies Press. https://www.nap.edu/catalog/10571/improving-the-design-of-the-scientists-and-engineers-statistical-data-system-sestat

National Science Foundation. undated. “Committee on Equal Opportunities in Science and Engineering (CEOSE),” webpage updating uncertain. https://www.nsf.gov/od/oia/activities/ceose/index.jsp

U.S. Government Printing Office. 2010. America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education and Science Reauthorization Act of 2010. https://www.govinfo.gov/content/pkg/PLAW-111publ358/html/PLAW-111publ358.htm

6. User Assessments of Transparent Reporting

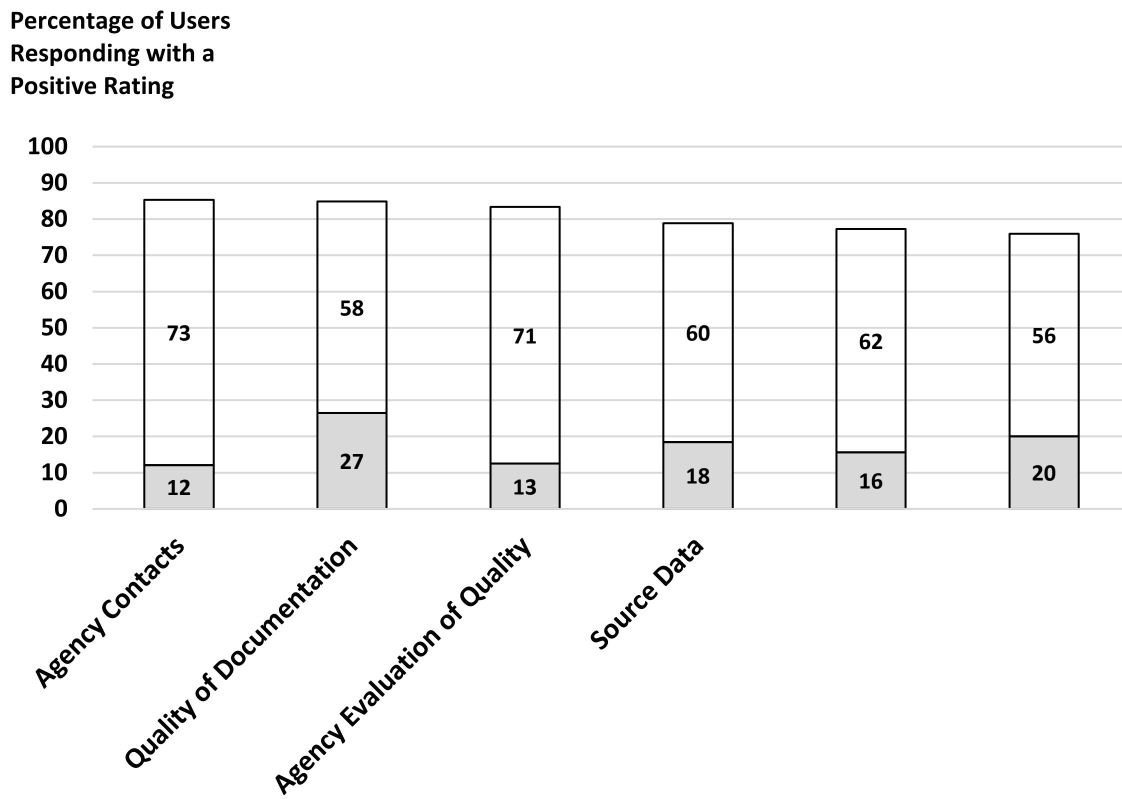
This chapter examines user assessments on documentation based on responses to 35 Lickert items on the customer survey on documentation. While the reviews of the case studies reported on the user responses to items soliciting unstructured, free-text responses, this chapter provides quantitative measures.

General Results. The indicator on which this section focuses is the percentage of users who have a positive rating of Satisfied or Very Satisfied (4 or 5 on a 5-point Lickert scale). Another indicator, the mean of user responses for an item, can be calculated once the Lickert scale is interpreted numerically. Both the percentages and the means are reported in Appendices 4 and 5. For the 35 items on documentation, the correlation between the two indicators—the percentages of positive ratings and the means—is 0.93. This high degree of correlation means that the information provided by either indicator is largely, albeit not completely, mirrored in the other.

To provide organization, the 35 Lickert item on documentation were grouped into six clusters of related items: Agency Contacts (7 items); Quality of Documentation (3 items); Agency Evaluation of Quality (2 items); Source Data (4 items); Data Integration (8 items); and How to Use the Product (11 items). By estimating intra-cluster means of the percentage of positive ratings, we can examine which clusters tend to have items that users consider to be relatively strong and which clusters users would value additional detail for transparent reporting.

[Figure follows on next page]

Figure 6.1. **User Assessments of Agency Documentation, by Cluster of Related Items**



Note: A positive rating is a response of "Somewhat satisfied" or "Very satisfied" on a 5-point Likert scale. A mean percentage of positive ratings for a cluster of related items is displayed above each box. The mean percentages of "Somewhat satisfied" and "Very satisfied" may not sum exactly to the mean percentage of positive ratings for a cluster.
Source: Transparent Reporting Project Customer Survey

From a user's perspective, areas of relative strength in the transparent reporting on data quality were Agency Contacts (obtaining information from agency staff) and general-level Quality of Documentation, each of which had means of 85 percent (rounded). The means for Agency Evaluation of Quality and for Source Data were 84 and 79 percent, respectively. Two clusters for which the percentages of users with positive ratings were relatively low were Data Integration and How to Use the Product, with means of 77 and 76 percent. Averaging across all 35 Licker items on transparent reporting, the overall mean was 80. That value happens to equal the mean for the 9 items on data quality and overall confidence in the data discussed elsewhere in the report.

Detailed Results for 35 Items on Documentation. In the report's introductory section, the user responses to 9 items on data quality overall confidence in the data were gathered from both subgroups of informed consumers and researchers. In contrast, not all 35 items on documentation were posed to both subgroups. Instead, 20 items were posed to both subgroups (46 users), 11 targeted items were posed only to researchers (38 users), and 4 targeted