American Economic Association

Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern
Author(s): Charles F. Manski
Source: *Journal of Economic Literature*, Vol. 53, No. 3 (SEPTEMBER 2015), pp. 631-653
Published by: American Economic Association
Stable URL: https://www.jstor.org/stable/43932407
Accessed: 10-08-2018 09:36 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

American Economic Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of Economic Literature*

JSTOR

Journal of Economic Literature 2015, 53(3), 631–653
http://dx.doi.org/10.1257/jel.53.3.631

Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern[†]

CHARLES F. MANSKI[*]

Federal statistical agencies in the United States and analogous agencies elsewhere commonly report official economic statistics as point estimates, without accompanying measures of error. Users of the statistics may incorrectly view them as error free or may incorrectly conjecture error magnitudes. This paper discusses strategies to mitigate misinterpretation of official statistics by communicating uncertainty to the public. Sampling error can be measured using established statistical principles. The challenge is to satisfactorily measure the various forms of nonsampling error. I find it useful to distinguish transitory statistical uncertainty, permanent statistical uncertainty, and conceptual uncertainty. I illustrate how each arises as the Bureau of Economic Analysis periodically revises GDP estimates, the Census Bureau generates household income statistics from surveys with nonresponse, and the Bureau of Labor Statistics seasonally adjusts employment statistics. I anchor my discussion of communication of uncertainty in the contribution of Oskar Morgenstern (1963a), who argued forcefully for agency publication of error estimates for official economic statistics. (*JEL* B22, C82, E23)

> *"Perhaps the greatest step forward that can be taken, even at short notice, is to insist that economic statistics be only published together with an estimate of their error. Even if only roughly estimated, this would produce a wholesome effect. Makers and users of economic statistics must both refrain from making claims and demands that cannot be supported scientifically. The publication of error estimates would have a profound influence on the whole situation."*
>
> *Morgenstern (1963a, pp. 304–05)*

[*] Department of Economics and Institute for Policy Research, Northwestern University. This research was supported in part by National Science Foundation grant SES-1129475. I am grateful to Bruce Spencer for numerous valuable discussions and comments, as well as continuing encouragement. I have benefited from helpful discussions with, and comments from, Robert Barbera, Nancy Bates, Dennis Fixler, Fane Lineback, Benjamin Reist, Misa Tanaka, David Wessel, Kirk White, Jonathan Wright, Tommy Wright, and several anonymous reviewers. I have also benefited from the opportunity to present this work at the April 2014 Bank of England Interdisciplinary Workshop on the Role of Uncertainty in Central Bank Policy and in seminars at the Federal Reserve Bank of New York and the U. S. Census Bureau.

[†] Go to http://dx.doi.org/10.1257/jel.53.3.631 to visit the article page and view author disclosure statement(s).

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

1. *Introduction*

Government agencies commonly communicate official economic statistics to the public in news releases that make little, if any, mention of uncertainty in the reported estimates. Technical publications documenting data and methods do acknowledge that official statistics are subject to error. These publications sometimes provide guidance on the magnitude of sampling errors, but they generally do not attempt to quantify nonsampling errors.

Some prominent American examples are the reporting of gross domestic product (GDP), employment, and household income by the US Department of Commerce, Bureau of Economic Analysis (BEA); US Department of Labor, Bureau of Labor Statistics (BLS); and the US Department of Commerce, Bureau of the Census (Census Bureau).[1]

BEA Reporting of GDP Growth: The BEA reports quarterly estimates of GDP growth. The agency initially reports an "advance" estimate based on incomplete data and then reports revisions one and two months later as further data become available. For example, a June 25, 2014 news release stated (BEA 2014):

> Real gross domestic product . . . decreased at an annual rate of 2.9 percent in the first quarter of 2014 according to the "third" estimate released by the Bureau of Economic Analysis. . . . The GDP estimate released today is based on more complete source data than were available for the "second" estimate issued last month. In the second estimate, real GDP was estimated to have decreased 1.0 percent.

Although this statement recognizes that estimates of GDP growth are subject to revision, BEA practice has been to report estimates without accompanying measures of potential error. While the news release observed that the second estimate was based on incomplete data, it did not acknowledge that the third estimate was likewise based on incomplete data and would be revised further when data collected annually rather than quarterly became available.

A recent publication by BEA staff explains the practice of reporting estimates without measures of error as a response to the presumed wishes of the users of GDP statistics. Fixler, Greenaway-McGrevy, and Grimm (2014) state:

> Given that BEA routinely revises its estimate's during the course of a year, one might ask why BEA produces point estimates of GDP instead of interval estimates. . . . Although interval estimates would inform users of the uncertainty surrounding the estimates, most users prefer point estimates, and so they are featured. However, BEA provides the information that enables an interested user to construct their own interval estimate (p. 2).

BLS Reporting on Employment: On the first Friday of each month, the BLS issues *The Employment Situation*, a news release reporting official employment statistics for the previous month. The reported unemployment rate is based on data on households sampled in the Current Population Survey (CPS). The statistic giving growth in nonfarm employment is based on data collected from employer establishments sampled in the Current Employment Statistics survey. For example, the BLS reported this on June 6, 2014 (BLS 2014b): "Total nonfarm payroll employment rose by 217,000 in May, and the unemployment rate was unchanged at 6.3 percent." Thus, the BLS reports employment statistics as point estimates, without measures of potential error.

[1] Throughout the text of the article, abbreviated names of government agencies are used after first mentions for the sake of conciseness. In the reference section, full names of agencies are used when government documents are cited.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

A technical note issued with the news release contains a section on *Reliability of the Estimates* that acknowledges the presence of errors.[2] The section describes the use of standard errors and confidence intervals to measure sampling error, providing some numerical illustrations.[3] It then turns to nonsampling errors, stating that they

> . . . can occur for many reasons, including the failure to sample a segment of the population, inability to obtain information for all respondents in the sample, inability or unwillingness of respondents to provide correct information on a timely basis, mistakes made by respondents, and errors made in the collection or processing of the data.

The technical note does not indicate the magnitudes of the nonsampling errors that may be present in the employment statistics.

Census Reporting of Income Statistics: Each year, the U.S. Census Bureau reports statistics on the income distribution based on data collected in the Annual Social and Economic (ASEC) supplement to the CPS. For example, in a news release issued September 12, 2012, the Census Bureau declared (Census Bureau 2012a): "The nation's official poverty rate in 2011 was 15.0 percent, with 46.2 million people in poverty. After three consecutive years of increases, neither the poverty rate nor the number of people in poverty were statistically different from the 2010 estimates."

Thus, the census release provided point estimates, acknowledged but did not quantify sampling error, and did not mention nonsampling error.

The Census Bureau's annual *Current Population Report* provides numerous statistics characterizing the income distribution and measures sampling error by providing 90 percent confidence intervals for various estimates (Census Bureau 2012b). However, the report does not measure nonsampling errors. A supplementary technical document describes some sources of nonsampling error, but it does not quantify them (Census Bureau 2012c).

Reporting official statistics as point estimates without adequate attention to error manifests a common tendency of policy analysts to project *incredible certitude* (Manski 2011, 2013). In the absence of agency guidance, some users of official statistics may naively assume that errors are small and inconsequential. Persons who understand that the statistics are subject to error must fend for themselves and conjecture the error magnitudes. Thus, users of official statistics may misinterpret the information that the statistics provide.

Why should misinterpretation be a concern? A broad reason is that governments, firms, and individuals use official statistics when making numerous important decisions. The quality of decisions may suffer if decisionmakers incorrectly take point estimates at face value or incorrectly conjecture error magnitudes. For example, a central bank monitoring statistics on GDP growth, inflation, and employment may misevaluate the status of the economy and consequently set inappropriate monetary policy.

Agencies could mitigate misinterpretation and better inform the nation about the state of the economy if they were to measure uncertainty in official statistics and communicate it regularly in their news releases and technical publications. Using established statistical principles, it should be straightforward for agencies to communicate sampling error in statistics based on survey data. A positive role model is the monthly release by

[2] See www.bls.gov/news.release/empsit.tn.htm.

[3] For example, the note accompanying the June 6, 2014 release states, "the confidence interval for the monthly change in total nonfarm employment from the establishment survey is on the order of plus or minus 90,000." It explains that this is a 90 percent interval.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

the Census Bureau of statistics on new residential home sales in the previous month. Expression of uncertainty is prominent in these releases. For example, the one for March 2014 begins this way (Census Bureau 2014):

> Sales of new single-family houses in March 2014 were at a seasonally adjusted annual rate of 384,000, according to estimates released jointly today by the US Census Bureau and the Department of Housing and Urban Development. This is 14.5 percent (±12.9%) below the revised February rate of 449,000 and is 13.3 percent (±9.9%) below the March 2013 estimate of 443,000.

The Explanatory Notes section of the release states, "All ranges given for percent changes are 90 percent confidence intervals and account only for sampling variability." This transparent expression of sampling uncertainty could easily be emulated when the BLS and Census Bureau report employment and income statistics in their news releases. Further precedent for communicating sampling uncertainty can be found in the news releases of private survey organizations, which routinely state a sampling "margin of error" when they report political polls and other survey findings.[4]

It is more challenging to measure nonsampling errors for official statistics. There are many sources of such errors and there has been no consensus about how to measure them. The BLS Technical Note quoted earlier lists five types of nonsampling error in employment statistics, but provides estimates of none of them. Numerous other typologies have been proposed in the literature on *total survey error*, which seeks to jointly characterize sampling and nonsampling error in statistics based on sample surveys.

Groves and Lyberg (2010) provide an informative historical synthesis and critique of research on total survey error, tracing the literature from Deming (1944) through over a hundred subsequent contributions. Groves and Lyberg view the concept as most useful in guiding survey design, stating:

> Total survey error is a concept that purports to describe statistical properties of survey estimates, incorporating a variety of error sources. For many within the field of survey methodology, it is the dominant paradigm. Survey designers can use total survey error as a planning criterion. Among a set of alternative designs, the design that gives the smallest total survey error (for a given fixed cost) should be chosen (pp. 849–50).

Mulry and Spencer (1991, 1993) report a rare effort to measure total survey error in an important setting, examining the many errors that may affect enumeration of the population in the decennial census.[5]

In principle, assessment of the implications of total survey error should require specification of a loss function that measures how error affects a particular use of an official statistic. Mulry and Spencer (1993) discuss various loss functions in the context of census enumeration. Nevertheless, rather than choose a loss function specific to the use of a statistic, the prevalent practice has been to view the frequentist mean square error (MSE) of an estimate as an appropriate omnibus measure of total survey error and to decompose the MSE into the sum of

[4] Groves and Lyberg (2010) call attention to the difference in practice between private and government survey organizations, writing (p. 864): "Indeed, it is unfortunate that press releases of surveys sponsored by many commercial media remind readers of sampling error . . . . , while press releases of U.S. federal statistical agencies do not usually provide such warnings."

[5] The Census Bureau has sought to characterize total survey error in some of its surveys by undertaking and publishing a periodic "Quality Profile." These profiles discuss nonsampling error at length, but they do not quantify them. See, for example, Census Bureau (1998).

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

variance and squared bias, measuring sampling and nonsampling error, respectively.

Statistical theory provides well-understood ways to measure variance, but has been relatively silent on the measurement of bias. Groves and Lyberg write:

> The total survey error format forces attention to both variance and bias terms. This is not a trivial value of the framework. Most statistical attention to surveys is on the variance terms—largely, we suspect, because that is where statistical estimation tools are best found. Biases deserve special attention because
>
> a. Their effect on inference varies as a function of the statistic; and
>
> b. They are studied most often through the use of gold-standard measurements, not internal replication common to error variance (p. 868).

The term "gold-standard measurements" refers to the availability of some external source of information assumed to provide an accurate measure of the statistic under study. The literature on total survey error has largely been silent on measurement of nonsampling error when, as is usually the case in practice, no gold-standard measurement is available. For this and other reasons, Groves and Lyberg conclude (p. 874): "The paradigm has been more successful as an intellectual framework than a unified statistical model of error properties of survey statistics."

Acknowledgment that measurement of nonsampling errors is challenging does not justify the prevailing practice of government agencies. Making good-faith efforts to measure both sampling and nonsampling error would be more informative than having agencies report official statistics as if they are truths. I am not certain who first urged agencies to measure and report error, but comments by Simon Kuznets (1948) about Department of Commerce practices in defining and reporting the national income accounts show that this pioneer of official economic statistics recognized the importance of the matter early on. Kuznets devoted a section of his article to "The Margins of Error" and wrote:

> What is urged here is more explicit and continuous consideration of margins of errors in economic statistics, particularly of the synthetic type involved in national income estimates. Neither the producers nor users of such estimates are inclined to devote much time to this problem; the former wish to arrive promptly at comprehensive and well articulated totals, the latter are eager to use the estimates to get light upon some question that seems to require urgent answer. The very fact that the estimates are cast in the form of unique series, not of ranges, is itself an invitation to treat them as firm results and tends to discourage questioning whether a total of x billion might not just as well read x + a or x − a. Consequently, users' attention should be called to the possibility of error, and experiments should be attempted on the ways in which the margins can be made known. It is not unlikely that, in the very process, means of actually improving the estimates themselves will be found (p. 178).

Soon after Kuznets, Oskar Morgenstern argued forcefully for agency publication of error estimates in the conclusion to his book *On the Accuracy of Economic Observations* (Morgenstern 1950, 1963a).[6] Summing up the lessons of his fifteen-chapter study of the many sources and manifestations of error, Morgenstern offered a damning indictment of the agencies that report official economic statistics to the public, writing:

> The process of improving data is an unending one. . . .There is, however, one area where definite action is possible, though it will take time before desirable results will become visible. That is to stop important government agencies, such as the President's Council of Economic Advisors, the various government departments,

[6] The first edition of the book was published in 1950 and went out of print in 1952. Morgenstern subsequently greatly expanded and rewrote the book, publishing it as the second edition in 1963. All mentions of the book in the present article refer to the second edition.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

> the Federal Reserve Board and other agencies, public and private, from presenting to the public economic statistics as if these were free from fault. Statements concerning month-to-month changes in the growth rate of the nation are nothing but absurd and even year-to-year comparisons are not much better. The same applies to variations in price levels, costs of living and many other items. It is for the economists to reject and criticize such statements which are devoid of all scientific value, but it is even more important for them not to participate in their fabrication (Morgenstern 1963a, p. 304).

He then called for regular publication of error estimates in the statement quoted at the beginning of this article, remarking that this is "Perhaps the greatest step forward that can be taken."

More recently, the Committee on National Statistics (CNSTAT) of the National Research Council, a pillar of the statistical establishment in the United States, has embraced communication of uncertainty in official statistics in its publication *Principles and Practices for a Federal Statistical Agency*, which recommends that agencies adhere to various good practices. Practice 4, titled "Openness About Sources and Limitations of the Data Provided," states this:

> A statistical agency should be open about the strengths and limitations of its data, taking as much care to understand and explain how its statistics may fall short of accuracy as it does to produce accurate data. Data releases from a statistical program should be accompanied by a full description of the purpose of the program; the methods and assumptions used for data collection, processing, and reporting; what is known and not known about the quality and relevance of the data; sufficient information for estimating variability in the data; appropriate methods for analysis that take account of variability and other sources of error; and the results of research on the methods and data (Citro and Straf 2013, p. 18).

Unfortunately, federal statistical agencies have not adhered to the practice recommended by Kuznets, Morgenstern, and CNSTAT. A BLS document on standards for information quality states that the agency applies the CNSTAT *Principles and Practices*, but it makes no explicit reference to nonsampling error (BLS 2014a). Thirty years ago, the Census Bureau convened a research conference on nonsampling error, at which Bureau Director John Keane announced initiation of an ambitious new "Nonsampling Error Exploration Program" (Census Bureau 1986, p. 186). However, the Census Bureau still does not measure and report such errors. The recent book-length document *Statistical Quality Standards* (Census Bureau 2013) contains a section on "Producing Measures and Indicators of Nonsampling Error," but the section does not adequately engage the measurement problem. In particular, the discussion of error due to survey nonresponse only calls for regular measurement of nonresponse rates, not for regular measurement of potential error due to nonresponse.[7]

With this background, the present article considers how agencies might constructively measure and communicate some potentially important forms of error in official statistics. In my discussion thus far, I have repeatedly invoked the traditional distinction between sampling and nonsampling error. However, considering how to structure the article, I have decided that this distinction yields an

[7] A mandate for measurement of nonresponse error appears only in Sub-Requirement D3-3.6, which states (p. 58): "Nonresponse bias analyses must be conducted when unit, item, or total quantity response rates for the total sample or important subpopulations fall below the following thresholds.

1. The threshold for unit response rates is 80 percent.
2. The threshold for item response rates of key items is 70 percent.
3. The threshold for total quantity response rates is 70 percent. (Thresholds 1 and 2 do not apply for surveys that use total quantity response rates.)"

Thus, there is no requirement for analysis of bias if unit response is above 80 percent and item response is above 70 percent. There is, moreover, no guidance on what would constitute an informative analysis of nonresponse bias.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

ineffective organizing principle for what lies ahead.

Survey statisticians find it appealing to distinguish sampling errors from all others because statistical theory addresses measurement of sampling error but does not, per se, provide a foundation for measurement of other errors. However, broad aggregation of nonsampling errors is too crude. Speaking at the 1986 census conference on nonsampling error, Director Keane recognized the aggregation problem well when he stated:

> The diversity of nonsampling error is notably impressive. It is impressive in its heterogeneity particularly when compared with the homogeneity of sampling error. No wonder sampling error is so much easier to measure and, therefore, so much more frequently measured. Perhaps the all-encompassing term itself—nonsampling error—conceptually spans too much for operational clarity and clean measurement (Bureau of the Census 1986, p. 185).

Considering the sources and implications of error from the perspective of users of economic statistics, rather than the perspective of statisticians, I think it essential to distinguish errors in measurement of well-defined concepts from uncertainty about the concepts that should be measured. I also think it useful to distinguish errors that diminish with time from ones that persist. To highlight these distinctions, I will separately discuss *transitory statistical uncertainty*, *permanent statistical uncertainty*, and *conceptual uncertainty*.

Transitory statistical uncertainty arises because data collection takes time. Agencies sometimes release a preliminary estimate of an official statistic in an early stage of data collection and revise the estimate as new data arrives. Hence, uncertainty may be substantial early on, but diminish as data accumulates. When new data increase sample size under a fixed sampling plan, sampling uncertainty diminishes with time. When new data yield observations on parts of a population that were not sampled earlier or yield more refined information on units sampled earlier, nonsampling uncertainty diminishes.

Permanent statistical uncertainty arises from incompleteness or inadequacy of data collection that does not diminish with time. Sampling uncertainty stemming from the finite ultimate size of survey samples is permanent. So is nonsampling uncertainty stemming from nonresponse and from the possibility that some respondents may provide inaccurate data.

Conceptual uncertainty arises from incomplete understanding of the information that official statistics provide about well-defined economic concepts or from lack of clarity in the concepts themselves. Thus, conceptual uncertainty concerns the interpretation of statistics, rather than their magnitudes. Survey statisticians find it particularly difficult to study conceptual uncertainty because knowledge of statistical theory does not suffice. A substantive understanding of the potential uses of official statistics is necessary.

To exemplify transitory statistical uncertainty, section 2 considers BEA initial measurement of GDP and the ensuing process of revising the estimate as new data arrives. Section 3 addresses permanent statistical uncertainty due to nonresponse in sample surveys, using nonresponse to income questions in the CPS to illustrate. Section 4 discusses the conceptual uncertainty inherent in BLS seasonal adjustment of employment statistics.

I could usefully discuss many more exemplars of the three types of uncertainty, such as temporal revision of inflation statistics, response errors in surveys, and conceptual uncertainty in definitions of poverty, unemployment, GDP, inflation, and other macroeconomic indices. However, this article would then quickly become book length. It would become more than book length if I were to attempt to review and assess the large academic research literatures that seek to shed light on the nature and implications

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

of data revisions, survey nonresponse and error, seasonal adjustment of time series, and alternative definitions of macroeconomic indices. Hence, my citation of research literature will be highly selective, the main aim being to present ideas about communication of uncertainty in official statistics.

I have subtitled the article "An Appraisal Fifty Years after Morgenstern" because I think that *On the Accuracy of Economic Observations* was an important contribution. Its virtues include a compelling basic message and many specific insights, albeit embedded within a daunting mass of detail that may understandably put off potential readers. Although Morgenstern achieved lasting fame among economists through his collaboration with John von Neumann in the development of game theory, he was close to a lone voice in his concern with error in official economic statistics. His book attracted notice and some controversy early on.[8] However, it has since faded from the attention of economists and appears not to have yielded any action by the statistical agencies. My impression is that present-day users of official statistics are largely unaware of the book's existence.[9] In what follows, I cite passages from the book, where appropriate, and also call attention to other contributions that warrant recognition by economists today.

2. *Transitory Uncertainty: Revisions in National Income Accounts*

2.1 *BEA Reporting of GDP*

As mentioned in the introduction, the BEA reports quarterly estimates of GDP. The BEA initially reports an *advance* estimate based on incomplete data available one month after the end of a quarter. It reports *second* and *third* estimates after two and three months, when additional data become available. In the summer of each year, when more extensive data collected on an annual rather than quarterly basis become available, BEA reports a *first annual* estimate and then revises it further in subsequent years. Every five years, the BEA reevaluates its operational definition of GDP and accordingly makes yet further revisions to the historical GDP record. I will not discuss the five-year revisions here because they concern conceptual rather than statistical uncertainty.

An article describing the measurement of GDP explains the reasons for revisions as follows:

> For the initial monthly estimates of quarterly GDP, data on about 25 percent of GDP, especially in the service sector, are not available, and so these sectors of the economy are estimated based on past trends and whatever related data are available. . . . The initial monthly estimates of quarterly GDP based on these extrapolations are revised as more complete data become available. . . . The successive revisions can be significant, but the initial estimates provide a snapshot of economic activity much like the first few seconds of a Polaroid photograph in which an image is fuzzy, but as the developing process continues, the details become clearer (Landefeld, Seskin, and Fraumeni 2008, p. 194).

How large do the revisions to the BEA estimates tend to be? Consider the month-to-month revisions for the first quarter of 2014. The news release quoted in the introduction stated that the second and

[8] An instance of controversy is an exchange between Raymond Bowman and Morgenstern in *The American Statistician*. Bowman was Assistant Director for Statistical Standards at the US Bureau of the Budget. Commenting on an article by Morgenstern in the magazine *Fortune* that summarized his book for a general audience (Morgenstern 1963b), Bowman (1964) acknowledged some of Morgenstern's concern with error in official statistics. However, Bowman downplayed the severity of the problem and expressed disappointment that Morgenstern had not done more to propose procedures that would reduce error. Morgenstern (1964) reacted strongly in defense of his work, restating his indictment of government practices that interpret published statistics as if they are essentially error free.

[9] The book does, however, continue to draw occasional attention from scholars who study the history of economic thought. See Kenessey (1997) and Boumans (2012).

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

third estimates were −1.0 and −2.9 percent, respectively. The advance estimate was +0.1 percent.[10] Thus, an advance estimate of small growth was revised to a third estimate of substantial decline two months later.

The revisions in first quarter 2014 were atypically large, but substantial revisions are common. Fixler, Greenaway-McGrevy, and Grimm (2011) report that the mean absolute revision (MAR) from the advance estimates to the second estimates of real GDP is 0.5 percentage points, from the advance estimates to the third estimates is 0.6 percentage points, and from the second to the third estimates is 0.3 percentage points. Considering the period 1983–2009, they report that the overall MARs to the advance, second, and third quarterly estimates (comparing these estimates with the latest available for the relevant quarter) were 1.31, 1.29, and 1.32 percentage points. Observing that the magnitude of the revisions tends not to diminish with time, despite the availability of more data when forming the second and third estimates, the authors state: "The lack of declines in the MARs of GDP in successive vintages of current quarterly estimates is a phenomenon that has been noted in nearly all of BEA's analyses of revisions" (p. 12). However, Fixler, Greenaway-McGrevy, and Grimm (2014) report that the magnitude of revisions does tend to diminish with time in the period 1993–2012.

2.2 *The Substantive Significance of Revisions*

While the magnitudes of revisions to BEA estimates of GDP are straightforward to compute, the substantive significance of the revisions is a matter of interpretation. Fixler, Greenaway-McGrevy, and Grimm (2011) remark at their beginning of their article that "Economic policy decisions should not need to be reconsidered in the light of revisions to GDP estimates, and policymakers should be able to rely on the early estimates as accurate indicators of the state of the economy" (p. 12). They provide an upbeat absolute perspective in the conclusion to their article, stating: "The estimates of GDP and GDI are accurate; the MARs for both measures are modestly above 1.0 percentage point" (p. 12). Landefeld, Seskin, and Fraumeni (2008) provide an upbeat comparative perspective, stating:

> In terms of international comparisons, the US national accounts meet or exceed internationally accepted levels of accuracy and comparability. The US real GDP estimates appear to be at least as accurate—based on a comparison of GDP revisions across countries—as the corresponding estimates from other major developed countries (p. 213).

Croushore (2011) offers a considerably more cautionary perspective, stating: "Until recently, macroeconomists assumed that data revisions were small and random and thus had no effect on structural modeling, policy analysis, or forecasting. But real-time research has shown that this assumption is false and that data revisions matter in many unexpected ways" (p. 73).

To illustrate, he gives a notable example:

> In January 2009, in the middle of the financial crisis that began in September 2008, the initial release of the national income accounts showed a decline in real gross domestic product (GDP) of 3.8 percent (at an annual rate) for the fourth quarter—a bad number for sure but not as bad as might be expected considering the damage caused by the financial meltdown. But one month later, the GDP growth rate was revised down by 2.4 percentage points, showing a decline in real GDP of 6.2 percent and confirming that the US economy was in the middle of the worst recession in over twenty-five years. The 2.4 percentage point downward revision from the initial release to the first revised number was the largest revision ever recorded for quarterly real GDP. Real-time data analysis of the history of revisions of real GDP shows us

[10] See www.bea.gov/newsreleases/national/gdp/2014/pdf/gdp1q14_adv.pdf for the news release giving the advance estimate.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

> that the largest revision came at a very inopportune moment (p. 73).

This example is an extreme case, but it provides a stark warning that BEA revisions to GDP may be quite large and occur at times when vital policy decisions must be made. Leaving aside the singular event of the financial crisis, I view the MARs reported by Fixler, Greenaway-McGrevy, and Grimm (2011) for the period 1983–2009 as too large to warrant their upbeat conclusion that BEA quarterly estimates of GDP are accurate. A statement made by Croushore (2011) seems more on the mark: "If monetary policy depends on short term growth rates, then clearly policy mistakes could be made if the central bank does not account for data uncertainty" (p. 77).

Related reservations were expressed by Morgenstern. Chapter 14 of Morgenstern (1963a) scrutinizes the construction of national income statistics. Section 2, on "Concepts of National Statistics," describes the substantial revisions that result from periodic conceptual changes in the definition of national income. Section 7, on "Absolute Size of the Estimates: Relative Changes and Revision," considers the statistical revisions made as new information becomes available. Summarizing the situation, he writes: "These observations, incidentally, should be viewed as casting serious doubts on the usefulness of national income figures for business cycle analysis" (p. 268). Presenting further evidence on the significance of revisions, Morgenstern (1964) presents a table showing that estimates of annual growth in national income in the United States in the period from 1947 to 1962 sometimes vary widely, depending on whether one uses initial estimates, last estimates, or other specified revisions to compute the growth rates.

2.3 *Measuring Uncertainty due to Revisions*

Informative communication of the transitory uncertainty of GDP estimates should be relatively easy to accomplish. The historical record of BEA revisions has been made accessible for study in two "real-time" data sets maintained by the Philadelphia and St. Louis Federal Reserve Banks. See Croushore (2011) for details regarding these data sets and similar ones for other nations. An early example of a real-time data set appears in Morgenstern (1963a, pp. 249–250) as table 23, which presents the revisions to estimates of national income made in the period 1947 through 1961.

Measurement of transitory uncertainty in GDP estimates is straightforward if one finds it credible to assume that the revision process is time stationary. Then historical estimates of the magnitudes of revisions can credibly be extrapolated to measure the uncertainty of future revisions. A particularly simple extrapolation would be to suppose that the historical overall MAR of 1.3 percentage points reported by Fixler, Greenaway-McGrevy, and Grimm (2011) will persist going forward. More broadly, it may be credible to suppose that the empirical distribution of revisions will persist going forward.

More refined measures of uncertainty can be developed by studying how the magnitude and direction of historical revisions have varied with the state of the economy. Such refinements may be important to the extent that the nature of revisions tends to vary over the business cycle (see Croushore 2011, section 2.2). If there is reason to think that the historical pattern of variation of revisions with the state of the economy will persist into the future, then it would be appropriate to measure the transitory uncertainty of future GDP estimates in a manner that conditions on the state of the economy.

A notable precedent for probabilistic communication of the transitory uncertainty in GDP estimates is the periodic release of fan charts by the Bank of England. I describe the British practice below.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

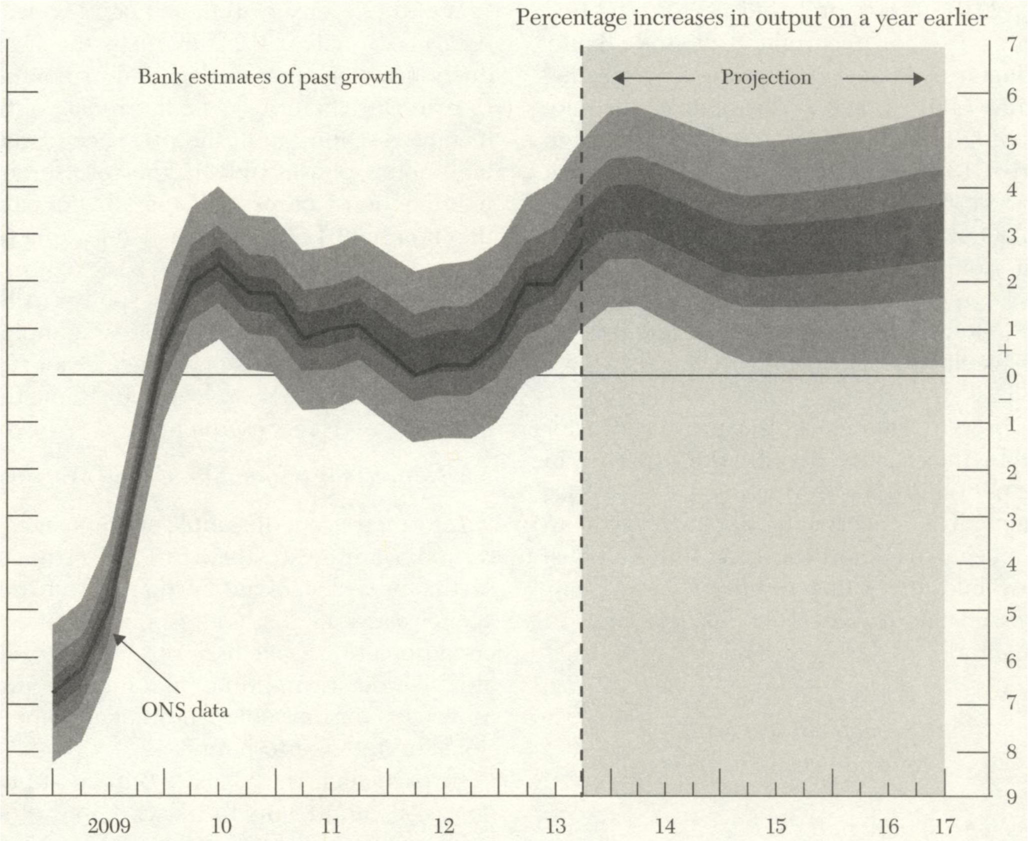

Figure 1. February 2014 UK GDP Fan Chart

Source: Bank of England

2.4 *Fan Chart Reports of GDP Growth by the Bank of England*

In the United Kingdom, the UK Office for National Statistics (ONS) reports quarterly estimates of GDP in a manner similar to the BEA, with no quantitative measurement of uncertainty, despite the fact that the estimates are revised regularly. For example, the April 2014 edition of the monthly *Economic Review* of the ONS describes the most recent GDP revisions as follows: "The Quarterly National Accounts left Gross Domestic Product (GDP) growth in the final quarter of 2013 unrevised at 0.7 [percent], but reduced annual growth in 2013 to 1.7 [percent], largely as a consequence of lower than previously estimated household expenditure" (ONS 2013, p. 1).

Unlike those in the United States, users of GDP estimates in the United Kingdom have ready access to a measure of transitory uncertainty in the fan charts reported by the Bank of England in its monthly *Inflation Report*. Figure 1 reproduces a fan chart for annual GDP growth in the *February 2014 Inflation Report* (Bank of England 2014). The part of the plot showing growth from

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

late 2013 on is a probabilistic forecast that expresses the uncertainty of the Bank's Monetary Policy Committee regarding future GDP growth. The part of the plot showing growth in the period 2009 through mid-2013 is a probabilistic forecast that expresses uncertainty regarding the revisions that ONS will henceforth make to its estimates of past GDP. The Bank explains as follows: "In the GDP fan chart, the distribution to the left of the vertical dashed line reflects the likelihood of revisions to the data over the past" (p. 7).

Observe that figure 1 expresses considerable uncertainty about GDP growth in the period 2010–13. Moreover, it expresses comparable uncertainty about growth in the recent past and the near future. Thus, the Bank judges that future ONS revisions to estimates of past GDP may be large in magnitude.

3. *Permanent Uncertainty: Nonresponse in Surveys*

3.1 *Nonresponse in the CPS*

Nonresponse is common in the surveys used to compute important official statistics. Unit and item nonresponse may make key data missing for substantial fractions of the persons sampled. For example, there is considerable nonresponse to the ASEC–CPS income questions used to produce the poverty statistics cited in the introduction. During the period 2002–12, 7 to 9 percent of the sampled households yielded no income data due to unit nonresponse and 41 to 47 percent of the interviewed households yielded incomplete income data due to item nonresponse (Manski 2014).[11]

Yet the Census Bureau news release (Census Bureau 2012a) cited in the introduction provides no explanation of how the Census Bureau forms a point estimate of the income distribution in the presence of such high nonresponse. Indeed, the release does not mention nonresponse at all. Similarly, the monthly BLS news release reporting the current unemployment rate (BLS 2014b) is silent on unit and item nonresponse to the CPS question used to estimate unemployment. These practices encourage readers of the census and BLS releases to think that nonresponse is inconsequential.

3.2 *Nonresponse Imputations and Weights*

Informed users of sample surveys such as the CPS are aware that, to form estimates in the presence of nonresponse, statistical agencies assume that nonresponse is random conditional on specified observed covariates. These assumptions are implemented as weights for unit nonresponse and imputations for item nonresponse.

In particular, the Census Bureau applies *hot-deck* imputations to the CPS and other surveys, describing the hot deck this way:

> This method assigns a missing value from a record with similar characteristics, which is the hot deck. Hot decks are defined by variables such as age, race, and sex. Other characteristics used in hot decks vary depending on the nature of the unanswered question. For instance, most labor force questions use age, race, sex, and occasionally another correlated labor force item such as full- or part-time status (Census Bureau 2006, chapter 9, p. 2).

[11] The ASEC questionnaire asks each member of the household about eighteen separate income components, ranging from earnings and pensions to dividends and public assistance. To determine total household income, the Census Bureau sums the responses obtained for these income components across the members of the household. The ASEC data shows a wide range of item nonresponse patterns, as households differ in the data they provide about the various income components. For example, nonresponse to the question asking for earnings on the primary job ranges from 17 to 19 percent over the ten-year period. Nonresponse to the question on interest income ranges from 23 to 27 percent over the period. Nonresponse on dividend income ranges from 9 to 12 percent.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

Thus, agency staff select a vector of covariates for which response is complete and determine the empirical distribution of the item of interest among sample members who have this covariate value and who report the item. A value for the item is imputed to a sample member with missing data by drawing a realization from the available empirical distribution.

The CPS documentation of hot-deck imputation offers no evidence that the method yields values for missing data that are close to the actual values. Another Census Bureau document describing the American Housing Survey is revealing. The document states:

> Some people refuse the interview or do not know the answers. When the entire interview is missing, other similar interviews represent the missing ones. . . . For most missing answers, an answer from a similar household is copied. The Census Bureau does not know how close the imputed values are to the actual values (Census Bureau 2011).

Indeed, lack of knowledge of the closeness of imputed values to actual ones is common.

Researchers have expressed varying views on the adequacy of census imputations. Considering imputation methodology in generality, some argue paternalistically that the staff of statistical agencies are better able to make decisions about treatment of missing data than are data users and, moreover, that imputation simplifies life for data users. A clear statement of this perspective appears in Meng (1994), who writes:

> A common technique for handling incomplete observations is to impute them before any substantive analysis. An obvious reason for the popularity of imputation, from a computational point of view, is that it allows users of the data to apply standard complete-data techniques directly. From an inferential point of view, perhaps, the most fundamental reason for imputation is that a data collector's assessment and information about the data, both observed and unobserved, can be incorporated into the imputations. In other words, imputation sensibly divides the tasks for analyzing incomplete data by assigning the difficult task of dealing with nonresponse mechanisms to those who are more capable of handling them, while allowing users to concentrate on their intended complete-data analyses (p. 538).

Many users of Census Bureau surveys appear to appreciate the simplification of data analysis that imputation yields. Considering census income imputations, Lillard, Smith, and Welch (1986) comment:

> Because much of what we think we know about income distributions and determinants of earnings is derived from studies of the censuses and CPSs, findings are sensitive to the treatment of non-reporters. Even so, the economic research community has ignored the problem, accepting census imputations as fully equivalent to reported values. In five leading economics journals during the last decade, over 100 articles used census or CPS data for studies of income; not one of them attempted to deal with the potential problems caused by non-random refusal to report. Part of this oversight may be attributed to tacit acceptance of census imputations and part may stem from inertia because many of the early releases of public-use tapes did not identify imputed values.

Lillard, Smith, and Welch express concern that empirical economists use census imputations as if they were real data. In particular, they question the maintained assumption that nonresponse is conditionally random. At the end of their extensive study of census imputation processes, they conclude that "we are left with a real concern about the accuracy of the income data that underlie a good deal of empirical economic research" (p. 505).

Reading Morgenstern, I find it striking that his entire dense book makes almost no mention of survey nonresponse. The only reference to it appears in chapter 13, on "Employment and Unemployment Statistics." A footnote describing the CPS notes the presence of some unit nonresponse, stating that about 1,500 households within a

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

sample of about 35,000 households are not interviewed (p. 219, footnote 3). Later in the chapter, Morgenstern observes that this "failure to obtain responses from a small proportion of designated sampling units" (p. 233) may be a source of error in labor-force statistics. I am aware of no mention of item nonresponse anywhere in the book, nor any mention of weighting or imputation as means of dealing with nonresponse.

Given that Morgenstern describes many sources of error in official statistics in considerable detail, the absence of attention to survey nonresponse calls for explanation. I conjecture that Morgenstern considered nonresponse to be a relatively small problem in his era, and so chose not to delve into it. The rate of CPS unit nonresponse that he cited was only about 4 percent. Lillard, Smith, and Welch (1986) observe that nonresponse has grown considerably with time—they report that item nonresponse to CPS income questions increased from 5.3 percent in 1948 to 26.6 percent in 1982 (p. 491, table 1). As mentioned earlier, item nonresponse has become even more pronounced recently (Manski 2014). When Morgenstern wrote his book in the early 1960s, he may not have anticipated that nonresponse would develop into the substantial problem that it has since become.

3.3 *Measuring Uncertainty due to Nonresponse*

Whereas imputations and weights embody the strong and often untenable assumption that nonresponse is conditionally random, econometric research on partial identification has shown how to measure uncertainty due to nonresponse without making any assumptions about the nature of the missing data. The idea, simply enough, is to contemplate all the values that the missing data might take. In doing so, the available data yield interval rather than point estimates of official statistics. See Manski (1989, 1994, 2003, 2007), Horowitz and Manski (1998, 2000), and Stoye (2010) inter alia. Econometricians have also shown how to form confidence intervals that jointly measure sampling and nonresponse error (e.g., Horowitz and Manski 2000; Imbens and Manski 2004; Stoye 2009). Thus, we know how to measure uncertainty for official statistics with survey nonresponse.

Determination of no-assumptions interval estimates is particularly easy when the statistic of interest is the mean or a quantile of an item. To obtain the lower (upper) bound of the interval, one supposes that all cases of nonresponse take the lowest (highest) logically possible value of the item and computes the resulting statistic. Thus, computation of the interval estimate just requires two extreme imputations of each case of missing data.[12]

A notable early precedent in the survey statistics literature was the Cochran, Mosteller, and Tukey (1954, pp. 274–82) consideration of the potential implications of nonresponse in the Kinsey survey of male sexual behavior. Considering an item with yes/no response options, they observed that the smallest (largest) possible value of the frequency of a yes response occurs if all missing data take the no (yes) value. However, the subsequent literature did not follow up on the idea. Summarizing his earlier work with Mosteller and Tukey, Cochran (1977) chose not to recommend further analysis of this type. Using the symbol W_2 to denote the nonresponse rate, he stated "The limits are distressingly wide unless W_2 is very small" (p. 362).

Indeed, no-assumptions interval estimates of statistics on the income distribution are distressingly wide, given the high rate of

[12] This procedure is valid for no-assumptions interval estimation of any statistic that respects stochastic dominance (see Manski 2003, section 1.3). Means and quantiles are leading examples. It is not valid for interval estimation of a spread statistic such as the variance or interquantile range of an item. Blundell et al. (2007) and Stoye (2010) derive no-assumptions intervals for spread statistics.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

nonresponse to the ASEC–CPS income questions. Manski (2014) used ASEC data collected in 2002–12 to form interval estimates of median household income and the fraction of families with income below the official poverty threshold in the years 2001–11. I reported one set of estimates that makes no assumptions about item nonresponse, but suppose that unit nonresponse is random. Another set of estimates makes no assumptions about either item or unit nonresponse.

The estimates show vividly that item nonresponse poses a huge potential problem for inference on the American income distribution, and that unit nonresponse exacerbates the problem. For example, the interval estimate for the family poverty rate in 2011 is [0.14, 0.34] if one makes no assumptions about item response, but assumes that unit nonresponse is random. The interval is [0.13, 0.39] if one drops the assumption that unit nonreponse is random.

I also used monthly CPS data to form interval estimates of the unemployment rate in March of 2002–12. While item nonresponse is a relatively minor source of error for the unemployment rate, unit nonresponse is highly consequential. When unit nonresponse is assumed random but nothing is assumed about item nonresponse, the interval estimate for the unemployment rate in March 2012 is [0.08, 0.09]. When the assumption of random unit nonresponse is dropped, the interval is [0.07, 0.16].

Observing the substantial width of no-assumption interval estimates of the income distribution and unemployment rate, one might judge that the Census Bureau and BLS should not bother to report them because the intervals are "too wide to be informative."[13] Nevertheless, I would argue that federal statistical agencies should report such intervals as they endeavor to communicate uncertainty to the public.

Whatever their width may be, interval estimates obtained without assumptions on nonresponse are valuable for several reasons. First, they are easy to compute and understand. Second, they are maximally credible in the sense that they express all logically possible values of the statistic of interest. I have long emphasized that beginning with the data alone establishes a domain of consensus among the users of statistics and serves to bound disagreements about their interpretation (Manski 1995, 2007). Third, no-assumptions interval estimates make explicit the fundamental role that assumptions play in inferential methods that yield tighter findings. Wide bounds reflect real data uncertainties that cannot be washed away by assumptions lacking credibility.

The above argument does not imply, of course, that statistical agencies should refrain from making assumptions about nonresponse. Interval estimates obtained with no assumptions may be excessively conservative if agency analysts have some understanding of the nature of nonresponse. There is a vast middle ground between interval estimation with no assumptions and point estimation assuming that nonresponse is conditionally random. Manski (2003, chapters 1 and 2; 2007, chapters 3 and 4) and Stoye (2010) characterize the identifying power of various such assumptions. Given the historical prominence of assumptions that nonresponse is random, agencies may find it appealing to consider assumptions formalizing the conjecture that missing data are not too different from observed data. Such assumptions may directly constrain the distribution of missing data, or may constrain the response propensities of persons with different outcomes. Manski (2014) suggests some alternatives that statistical agencies may want to consider.

[13] I place the phrase "too wide to be informative" in quotation marks because I have often heard colleagues use this or a similar phrase when discussing interval estimates that make no assumptions about the nature of nonresponse.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

It is unlikely that any one middle-ground assumption will be appropriate in all settings. Hence, I will not propose adoption of any particular assumption for reporting of official statistics. A particularly simple broad idea is to use available information to conjecture an interval within which each missing data item lies. Thus, one would impute an interval of values when nonresponse occurs. This idea encompasses the two poles of traditional point imputation, in which the imputed interval contains only one value, and the no-assumption interval, in which the imputed interval contains all logically possible values of the item. Middle-ground assumptions impute intervals that are larger than a point but smaller than all logically possible values.[14]

3.4 *Reducing Nonresponse*

While the focus of this article is communication of uncertainty in official statistics, I would be remiss not to mention the efforts that survey designers might make to reduce nonresponse. Research on survey methodology has long sought to understand how survey mode (e.g., face-to-face, telephone, internet), participation incentives, questionnaire length, and the wording of questions affect unit and item nonresponse. Research on total survey error has encouraged survey planners to assess the cost-effectiveness of alternative designs. Spencer (1985) goes further and proposes benefit–cost analysis of alternative designs.

When considering nonresponse, an instructive cost-effectiveness exercise begins with a specified design and asks whether a marginal increase in survey budget should be used to increase the size of the sample or reduce nonresponse among existing sample members. Whereas the former use of budget reduces sampling uncertainty alone, the latter may reduce sampling and nonsampling uncertainty. Cochran, Mosteller, and Tukey (1954) reported an exercise of this type in the context of the Kinsey survey with no assumptions on nonresponse, using mean square error as the measure of total survey error. Horowitz and Manski (1998) reported a similar exercise in the context of CPS estimation of the national unemployment rate with no assumptions on nonresponse, using the confidence interval on the unemployment rate as the measure of total survey error. Both exercises suggest criteria for deciding when efforts to reduce nonresponse may be cost-effective.

In some cases, a feasible alternative to reduction of nonresponse may be to obtain administrative data that are informative about missing survey items. In particular, consider nonresponse to income questions in the CPS. The *1973 CPS–IRS–SSA Exact Match Study*, described in Kilss and Scheuren (1978), was an ambitious joint effort of the Census Bureau and the Social Security Administration to match the sample of respondents to the March Income Supplement to the CPS (now known as the ASEC Supplement) to their Social Security benefit and earning records, as well as to some information in their federal income tax returns. In principle, match studies of this type can be informative not only about CPS nonresponse, but also about the accuracy of the data provided by CPS sample members who respond to the income questions. Kilss and Scheuren (1978) write:

> The 1973 study was designed with a great number of specific goals in mind. . . . The

[14] The interval imputation described here differs fundamentally from the "multiple imputation" studied in Rubin (1987). The latter assumes that data are missing at random conditional on specified covariates. In a manner similar to the hot-deck method, one determines the empirical distribution of the item among sample members who have this covariate value and imputes a value to each sample member with missing data by drawing a realization at random from the empirical distribution. The word "multiple" refers to carrying out this process multiple times, creating multiple artificial data sets. The aim is to appropriately characterize the sampling uncertainty resulting from imputation.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

> primary interest of the Bureau of the Census, for example, was to evaluate and, potentially, to find ways of improving upon the procedures it employs in carrying out the Current Population Survey. . . It is important to add that the Exact Match Study was also looked upon as an intermediate step in the construction of corrected personal income-size distributions of the US population (p. 15).

I find the latter goal intriguing. It suggests the possibility of using SSA and IRS records to revise the estimates of the income distributions that are presently based purely on CPS responses. If the ASEC–CPS sample were matched with their SSA and IRS records each year, the result could be a data-revision process that would transform some of what is now permanent uncertainty in estimation of the income distribution into transitory uncertainty. Although it appears that the Census Bureau did not pursue this idea following the 1973 study, it warrants renewed consideration today. A recent study by Hokayem, Bollinger, and Ziliak (2014) explores what may be possible when CPS sample members are matched to their Social Security earnings records.

4. *Conceptual Uncertainty: Seasonal Adjustment of Official Statistics*

I wrote in the introduction that conceptual uncertainty arises from incomplete understanding of the information that official statistics provide about well-defined economic concepts or from lack of clarity in the concepts themselves. Section 2 mentioned one example, this being that the BEA reevaluates its operational definition of GDP every five years and revises the historical GDP record accordingly. Morgenstern (1963a) discussed difficulties in defining national income in chapter 14, section 2. Earlier, Kuznets (1948) offered a broad appraisal of the many conceptual issues.

Another example arises in the BLS measurement of unemployment. The BLS classifies CPS respondents as unemployed if they "do not have a job, have actively looked for work in the prior 4 weeks, and are currently available for work."[15] Responses to a sequence of CPS questions are used to determine whether a person has "actively looked for work in the prior 4 weeks" and is "currently available for work." The notions of actively looking for work and being currently available for work are inherently vague to some degree, so there is resulting uncertainty about how unemployment statistics should be interpreted.

4.1 *Seasonal Adjustment in Principle and Practice*

A particularly troubling conceptual uncertainty arises in the conventional practice of seasonally adjusting official statistics, including quarterly GDP estimates and monthly unemployment rates. Viewed from a sufficiently high altitude, the purpose of seasonal adjustment appears straightforward to explain. However, it is much less clear from ground level how one should actually perform seasonal adjustment.

The BLS explains seasonal adjustment of employment statistics this way (BLS 2001):

> What is seasonal adjustment? Seasonal adjustment is a statistical technique that attempts to measure and remove the influences of predictable seasonal patterns to reveal how employment and unemployment change from month to month. Over the course of a year, the size of the labor force, the levels of employment and unemployment, and other measures of labor market activity undergo fluctuations due to seasonal events including changes in weather, harvests, major holidays, and school schedules. Because these seasonal events follow a more or less regular pattern each year, their influence on statistical trends can be eliminated by seasonally adjusting the statistics

[15] See www.bls.gov/cps/cps_htgm.htm#unemployed.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

> from month to month. These seasonal adjustments make it easier to observe the cyclical, underlying trend, and other nonseasonal movements in the series.

The explanation is heuristically appealing, but it does not specify how, in practice, one may "remove the influences of predictable seasonal patterns."

Views on appropriate ways to perform seasonal adjustment have long varied. See, for example, the exchange between Granger (1978) and Sims (1978), as well as the recent contribution of Wright (2013). The absence of consensus about the right way to perform seasonal adjustment has also played a role in policy debates. A fascinating instance appears in a 1961 letter to the Editor of the *New York Times* written by Paul Samuelson, who served as an economic advisor to President Kennedy.

Samuelson (1961) compares the time series of unemployment that emerge with two methods of seasonally adjusting the unemployment rate, one being the then official BLS method and the other called the "residual method." He writes:

> The American economy has been in a fairly vigorous rise ever since last February. Everything seems to be improving: production, income, wages and profits. The one flaw in the picture has been the apparent failure of unemployment to improve. At last report (October) 6.8 percent of our civilian labor force was stated to be unemployed on a seasonally corrected basis, which suggests no improvement in the level that we have been experiencing for the previous eleven months of recession and recovery. How can we square this with the fact that the Federal Reserve Board's index of physical production has risen from its February low by about 10 percent? . . . I should like to suggest that the answer is a simple one. . . . . While correct in presenting the ground level of our joblessness and its long-term trends, the official method of seasonally correcting raw unemployment statistics to arrive at the best estimate of unemployment is open to reasonable questioning. . . . If you replace the official method by an alternative technical method sponsored by a number of academic and other experts (the so-called "residual method". . .), you find that unemployment has been falling steadily from its February peak of 7.2 per cent of the labor force to the October level of 6.4 per cent.

Thus, Samuelson points out that use of the residual method to seasonally adjust the unemployment rate yields a distinctly more positive conclusion regarding the success of the Kennedy economic policy than does use of the official BLS method.

Today the BLS uses the X-12-ARIMA method, developed by the Census Bureau and described in Findley et al. (1998). The X-12 method, along with its predecessor X-11 and successor X-13, may be a sophisticated and successful algorithm for seasonal adjustment. Or it may be an unfathomable black box containing a complex set of statistical operations that lack economic foundation. Wright (2013) eloquently expresses the difficulty of understanding X-12, writing:

> Most academics treat seasonal adjustment as a very mundane job, rumored to be undertaken by hobbits living in holes in the ground. I believe that this is a terrible mistake, but one in which the statistical agencies share at least a little of the blame. Statistical agencies emphasize SA data (and in some cases don't even publish NSA data), and while they generally document their seasonal adjustment process thoroughly, it is not always done in a way that facilitates replication, or encourages entry into this research area (p. 67).

Wright's remark that statistical agencies sometimes do not publish nonseasonally adjusted (NSA) data refers particularly to a decision of the BEA to stop publication of nonadjusted GDP estimates. He comments on this as follows: "It is very unfortunate that for the most basic measure of economic activity in the largest country in the world, researchers are effectively prevented from

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

evaluating any difficulties associated with seasonal adjustment" (p. 79).[16]

Understanding the practice of seasonal adjustment matters because, as Wright states, "Seasonal adjustment is extraordinarily consequential" (p. 65). He gives this example concerning BLS reporting of estimates of nonfarm payrolls: "In monthly change, the average absolute difference between the SA and NSA number is 660,000, which dwarfs the normal month-over-month variation in the SA data. All this implies that we should think very carefully about how seasonal adjustment is done" (p. 65).

Fifty years earlier Morgenstern (1963a) similarly warned against overinterpretation of month-to-month movements in seasonally adjusted unemployment rates. He wrote:

> It is not uncommon, indeed it is frequent, to find the government making strong statements about developments in unemployment over periods as short as *one* (!) month. The nation's largest or most important newspapers play up a "drop" in the unemployment rate, say from 5.8% to 5.5% as a highly significant event and the Secretary of Labor will not hesitate to make speeches on that occasion. All this is done on the basis of "seasonal correction" and dealing with figures given to four "significant" digits. It is, of course, clear that statements of this kind are completely devoid of the meaning attributed to them (p. 239).

4.2 *Measurement of Uncertainty Associated with Seasonal Adjustment*

There presently exists no clearly appropriate way to measure the conceptual uncertainty associated with seasonal adjustment. The Census Bureau's X-12 is an algorithm, not a method based on a well-specified dynamic theory of the economy. Hence, it is not obvious how to evaluate the extent to which it accomplishes the objective of removing the influences of predictable seasonal patterns. One might perhaps juxtapose X-12 with other seemingly reasonable algorithms, perform seasonal adjustment with each one, and view the range of resulting estimates as a measure of conceptual uncertainty.

More principled ways to evaluate uncertainty may open up if statistical agencies were to use a seasonal adjustment method derived from a well-specified model of the economy. One could then assess the sensitivity of seasonally adjusted estimates to variation in the parameters and the basic structure of the model.

A more radical departure from present practice would be to abandon seasonal adjustment and leave it to the users of official statistics to interpret unadjusted statistics as they choose. Publication of unadjusted statistics should be particularly valuable to users who want to make year-to-year rather than month-to-month comparisons of statistics. Suppose, for example, that one wants to compare unemployment in March 2013 and March 2014. It is arguably more reasonable to compare the unadjusted estimates for these months than to compare the seasonally adjusted estimates. Comparison of unadjusted estimates for the same month each year (March in this case) sensibly removes the "influences of predictable seasonal patterns" that the BLS noted in its

[16] Curious as to why the BEA stopped publication of NSA GDP estimates, I posed the question to Dennis Fixler, Chief Statistician of the BEA, in an email message. He replied (email, May 12, 2014):

> "regarding the comment made by Jonathan Wright, we had to drop the nonseasonally adjusted BEA data for budget reasons. Unfortunately, our budget is quite tight and the prospects for increases are not great. Consequently, there are no plans to reinstitute these data."

In a later message, he elaborated on the reasons why publication of nonadjusted GDP estimates is costly, stating (email, May 15, 2014):

> "Much of BEA's source data for GDP comes from the source data agencies already seasonally adjusted (we don't get the unadjusted data), and separate collections were needed for the unadjusted estimates, and a good deal of work was needed to produce the nonseasonally adjusted estimates. Also, until the third annual revision vintage estimates, the GDP estimates contain some judgmental inputs that are inherently seasonally adjusted."

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

2001 document. Moreover, it compares data actually collected in the two months of interest. In contrast, the seasonally adjusted estimates for March 2013 and March 2014 are composed of data collected not only in these months, but over a lengthy prior period.

5. *Conclusion*

Over fifty years ago, Oskar Morgenstern urgently argued for regular measurement of error in official economic statistics. He made a considerable effort to change the prevailing practices of federal statistical agencies in the United States, writing two editions of *On the Accuracy of Economic Observations* and articles summarizing the book. He was well-placed to influence the status quo, being famous for his contribution to game theory and situated prominently at Princeton. Yet his efforts did not bear fruit. Nor have agencies adhered to the Committee on National Statistics call for "Openness about Sources and Limitations of the Data Provided" in *Principles and Practices for a Federal Statistical Agency*.

Why is it that federal statistical agencies still do so little to communicate uncertainty in official statistics? I am unaware of any valid professional reason that would explain the failure of the BLS and census to report measures of sampling error in their news releases of employment and income statistics. These agencies know how to measure sampling error and they do so in their technical documents. Private survey organizations routinely state a sampling "margin of error" in their news releases.

Considering nonsampling error, one might fault Morgenstern, a sharp critic of agency practices, for not proposing specific constructive ways to measure such errors. Nor does the CNSTAT document propose measurement ideas. Yet, I do not think that this explains the continuing reluctance of statistical agencies to quantify nonsampling error.

This article has suggested several specific actions that agencies could take now, implementing ideas that have already been developed. Emulating the Bank of England, the BEA could prepare and publish fan charts to depict uncertainty in the GDP revision process. The BLS and Census Bureau could use no-assumptions interval estimates or more informative interval imputations to communicate uncertainty generated by survey nonresponse. Agencies that seasonally adjust official statistics could make available unadjusted statistics.

Going further, agency administrators could task their research staffs to develop measures of nonsampling error, perhaps in collaboration with external econometricians and statisticians. Federal statistical agencies have previously allocated resources to develop the methods now used to weight, impute, and seasonally adjust official statistics. They could similarly develop ways to measure nonsampling error. Generation of useful measurement methods will require agencies to exercise judgment, but this is no reason to deter them from making the effort. Agencies already make judgments when they extrapolate trends to construct advance GDP estimates, when they use weights and imputations to deal with survey nonresponse, and when they seasonally adjust a multitude of statistics. They implicitly judge nonsampling error to be inconsequential when they choose not to quantify it. They should reject this flawed implicit judgment and use their professional expertise to develop useful measures of nonsampling error.

While I cannot conjure a valid professional explanation for the status quo, I do see a possible political explanation. I have elsewhere observed that policymakers and the public appear to want to receive policy analysis that expresses certitude, even though the certitude may lack credibility (Manski 2011; 2013, chapter 1). I was referring then to incentives for incredible certitude facing

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

researchers and government agencies that predict policy outcomes.[17] Federal statistical agencies may similarly perceive an incentive to express incredible certitude about the state of the economy when they publish official economic statistics.

Morgenstern (1963a) comments cogently on the political incentives facing statistical agencies when he writes:

> Finally, we mention a serious organizational difficulty in discussing and criticizing statistics. These are virtually always produced by large organizations, government or private; and these organizations are frequently mutually dependent upon each other in order to function normally. Often one office cannot publicly raise questions about the work of another, even when it suspects the quality of the work, since this might adversely affect bureaucratic-diplomatic relations between the two and the flow of information from one office to another might be hampered. A marked esprit de corps prevails. All offices must try to impress the public with the quality of their work. Should too many doubts be raised, financial support from Congress or other sources may not be forthcoming. More than once has it happened that Congressional appropriations were endangered when it was suspected that government statistics might not be 100 percent accurate. It is natural, therefore, that various offices will defend the quality of their work even to an unreasonable degree (p. 11).

Later, Morgenstern cites these political pressures as a reason that agencies publish official statistics with unjustifiable detail, a phenomenon that he calls "Chapter III: Specious Accuracy."

Not having the persuasive power of Congressional appropriations, I can only say that federal statistical agencies would better inform policymakers and the public if they were to measure and communicate important uncertainties in official statistics. Should agencies take the task seriously, I think it likely that they will want to develop separate strategies for communication of uncertainty in news releases and in technical documentation of official statistics.

News releases are brief and are aimed at a broad audience. Hence, they have only a limited ability to convey nuance. It will be challenging for agencies to write releases that effectively quantify both sampling and nonsampling error, but it is important that they do so. Agencies have more scope for communication of uncertainty in their technical documentation of official statistics. The possibilities grow as the norm in documentation moves away from publication of periodic reports and towards development of web-based software that enables the users of official statistics to flexibly access and study the available data.

An open question is how communication of uncertainty would affect policy making and private decision making. We now have little understanding of the ways that users of official statistics interpret them. Some may mistakenly take the statistics at face value. Others may conjecture that the statistics are prone to errors of varying directions and magnitudes. We know essentially nothing about how decision making would change if statistical agencies were to communicate uncertainty regularly and transparently.

Consider, for example, the use by policymakers of early estimates of GDP growth. An anonymous reviewer of this article posed a set of pertinent questions, which I paraphrase as follows:

(1) Are policy decisions ever made on the basis of early estimates or is it an

[17] For example, a core function of the Congressional Budget Office is to make ten-year point predictions of the budgetary impact of pending legislation. These predictions, called scores, are conveyed in letters that the director writes to leaders of Congress and chairs of Congressional committees. They are not accompanied by any measures of uncertainty, even though legislation often proposes complex changes to federal law, whose budgetary implications must be difficult to foresee.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

accepted fact that early numbers are such bad predictors that policymakers avoid them altogether?

(2) What is the cost of incorrect policy decisions made based on early numbers?

(3) Would policy decisions be different in the absence of early numbers?

(4) Would policy decisions be different had BEA reported a measure of uncertainty?

Analogous questions should be asked about policy making with household income statistics based on surveys with large nonresponse, with seasonally adjusted employment statistics, and with many other official statistics that now suppress measurement of uncertainty. I urge behavioral and social scientists to initiate empirical studies that would shed light on these matters.

References

Bank of England. 2014. *Inflation Report Fan Charts February 2014*. http://www.bankofengland.co.uk/publications/Documents/inflationreport/2014/ir14febfc.pdf.

Blundell, Richard, Amanda Gosling, Hidehiko Ichimura, and Costas Meghir. 2007. "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds." *Econometrica* 75 (2): 323–63.

Boumans, Marcel. 2012. "Observations in a Hostile Environment: Morgenstern on the Accuracy of Economic Observations." *History of Political Economy* 44 (Supplement 1): 114–36.

Bowman, Raymond T. 1964. "Comments on 'Qui Numerare Incipit Errare Incipit' by Oskar Morgenstern." *American Statistician* 18 (3): 10–20.

Citro, Constance F., and Miron L. Straf, eds. 2013. *Principles and Practices for a Federal Statistical Agency*, Fifth edition. Washington, DC: National Academies Press.

Cochran, William G. 1977. *Sampling Techniques*, Third edition. New York: John Wiley and Sons.

Cochran, William G., Frederick Mosteller, and John W. Tukey. 1954. *Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male*. Washington, DC: American Statistical Association.

Croushore, Dean. 2011. "Frontiers of Real-Time Data Analysis." *Journal of Economic Literature* 49 (1): 72–100.

Deming, W. Edwards. 1944. "On Errors in Surveys." *American Sociological Review* 9 (4): 359–69.

Findley, David F., Brian C. Monsell, William R. Bell, Mark C. Otto, and Bor-Chung Chen. 1998. "New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program." *Journal of Business and Economic Statistics* 16 (2): 127–52.

Fixler, Dennis J., Ryan Greenaway-McGrevy, and Bruce T. Grimm. 2011. "Revisions to GDP, GDI, and Their Major Components." *Survey of Current Business* 91 (7): 9–31.

Fixler, Dennis J., Ryan Greenaway-McGrevy, and Bruce T. Grimm. 2014. "The Revisions to GDP, GDI, and Their Major Components." *Survey of Current Business* 94 (8): 1–23.

Granger, Clive W. J. 1978. "Seasonality: Causation, Interpretation, and Implications." In *Seasonal Analysis of Economic Time Series*, edited by Arnold Zellner, 33–56. Cambridge, MA: National Bureau of Economic Research.

Groves, Robert M., and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74 (5): 849–79.

Hokayem, Charles, Christopher R. Bollinger, and James P. Ziliak. 2014. "The Role of CPS Nonresponse on the Level and Trend in Poverty." University of Kentucky Center for Poverty Research Discussion Paper 2014-05.

Horowitz, Joel L., and Charles F. Manski. 1998. "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations." *Journal of Econometrics* 84 (1): 37–58.

Horowitz, Joel L., and Charles F. Manski. 2000. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association* 95 (449): 77–84.

Imbens, Guido W., and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72 (6): 1845–57.

Kenessey, Zoltan. 1997. "A Perspective on the Accuracy of Economic Observations." *International Statistical Review* 65 (2): 247–59.

Kilss, Beth, and Frederick J. Scheuren. 1978. "The 1973 CPS–IRS–SSA Exact Match Study." *Social Security Bulletin* 41 (10): 14–22.

Kuznets, Simon. 1948. "Discussion of the New Department of Commerce Income Series." *Review of Economics and Statistics* 30 (3): 151–79.

Landesfeld, J. Steven, Eugene P. Seskin, and Barbara M. Fraumeni. 2008. "Taking the Pulse of the Economy: Measuring GDP." *Journal of Economic Perspectives* 22 (2): 193–216.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation." *Journal of Political Economy* 94 (3 Part 1): 489–506.

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources* 24 (3): 343–60.

Manski, Charles F. 1994. "The Selection Problem." In *Advances in Econometrics, Sixth World Congress, Volume 1*, edited by Christopher A. Sims, 143–70.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms

Cambridge and New York: Cambridge University Press.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA and London: Harvard University Press.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York: Springer.

Manski, Charles F. 2007. *Identification for Prediction and Decision*. Cambridge, MA. and London: Harvard University Press.

Manski, Charles F. 2011. "Policy Analysis with Incredible Certitude." *Economic Journal* 121 (554): F261–89.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Cambridge: Harvard University Press.

Manski, Charles F. 2014. "Credible Interval Estimates for Official Statistics with Survey Nonresponse." Northwestern University Department of Economics.

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9 (4): 538–58.

Morgenstern, Oskar. 1950. *On the Accuracy of Economic Observations*. Princeton and London: Princeton University Press.

Morgenstern, Oskar. 1963a. *On the Accuracy of Economic Observations*, Second edition. Princeton and London: Princeton University Press.

Morgenstern, Oskar. 1963b. "Qui Numerare Incipit Errare Incipit." *Fortune* October.

Morgenstern, Oskar. 1964. "Fide Sed Ante Vide: Remarks to Mr. R. T. Bowman's 'Comments'." *American Statistician* 18 (4): 15–25.

Mulry, Mary H., and Bruce D. Spencer. 1991. "Total Error in PES Estimates of Population." *Journal of the American Statistical Association* 86 (416): 839–55.

Mulry, Mary H., and Bruce D. Spencer. 1993. "Accuracy of the 1990 Census and Undercount Adjustments." *Journal of the American Statistical Association* 88 (423): 1080–91.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Samuelson, Paul A. 1961. "Estimating Unemployment: Samuelson Questions Method of Seasonally Adjusting Figures." *The New York Times* November 12.

Sims, Christopher A. 1978. "Comments on 'Seasonality: Causation, Interpretation, and Implications' by Clive W. J. Granger." In *Seasonal Analysis of Economic Time Series*, edited by Arnold Zellner, 47–49. Cambridge, MA: National Bureau of Economic Research.

Spencer, Bruce D. 1985. "Optimal Data Quality." *Journal of the American Statistical Association* 80 (391): 564–73.

Stoye, Jörg. 2009. "More on Confidence Intervals for Partially Identified Parameters." *Econometrica* 77 (4): 1299–1315.

Stoye, Jörg. 2010. "Partial Identification of Spread Parameters." *Quantitative Economics* 1 (2): 323–57.

UK Office for National Statistics. 2013. *Economic Review, April 2014*. http://www.ons.gov.uk/ons/dcp171766_358477.pdf.

US Department of Commerce, Bureau of Economic Analysis. 2014. "News Release: Gross Domestic Product: First Quarter 2014 (Third Estimate)." http://www.bea.gov/newsreleases/national/gdp/2014/pdf/gdp1q14_3rd.pdf.

US Department of Commerce, Bureau of the Census. 1986. *Second Annual Research Conference Proceedings*.

US Department of Commerce, Bureau of the Census. 1998. "SIPP Quality Profile 1998." Survey of Income and Program Participation Working Paper 230.

US Department of Commerce, Bureau of the Census. 2006. "Design and Methodology: Current Population Survey." http://www.census.gov/prod/2006pubs/tp-66.pdf.

US Department of Commerce, Bureau of the Census. 2011. *American Housing Survey for the United States: 2009*. Washington, DC: US Government Printing Office.

US Department of Commerce, Bureau of the Census. 2012a. "Income, Poverty and Health Insurance Coverage in the United States: 2011." http://www.census.gov/newsroom/releases/archives/income_wealth/cb12-172.html.

US Department of Commerce, Bureau of the Census. 2012b. "Income, Poverty, and Health Insurance Coverage in the United States: 2011." https://www.census.gov/prod/2012pubs/p60-243.pdf.

US Department of Commerce, Bureau of the Census. 2012c. "Source and Accuracy of Estimates for *Income, Poverty, and Health Insurance Coverage in the United States: 2011*." http://www.census.gov/hhes/www/p60_243sa.pdf.

US Department of Commerce, Bureau of the Census. 2013. "Statistical Quality Standards." http://www.census.gov/quality/standards/Quality_Standards.pdf.

US Department of Commerce, Bureau of the Census. 2014. "New Residential Sales in March 2014." http://www.census.gov/construction/nrs/pdf/newressales.pdf.

US Department of Labor, Bureau of Labor Statistics. 2001. "Labor Force Statistics from the Current Population Survey." http://www.bls.gov/cps/seasfaq.htm.

US Department of Labor, Bureau of Labor Statistics. 2014a. "BLS Guidelines for Informing Users of Information Quality and Methodology." http://www.bls.gov/bls/quality.htm.

US Department of Labor, Bureau of Labor Statistics. 2014b. "Employment Situation News Release." http://www.bls.gov/news.release/archives/empsit_06062014.htm.

Wright, Jonathan H. 2013. "Unseasonal Seasonals?" *Brookings Papers on Economic Activity* Fall: 65–110.

This content downloaded from 158.109.174.204 on Fri, 10 Aug 2018 09:36:26 UTC
All use subject to https://about.jstor.org/terms