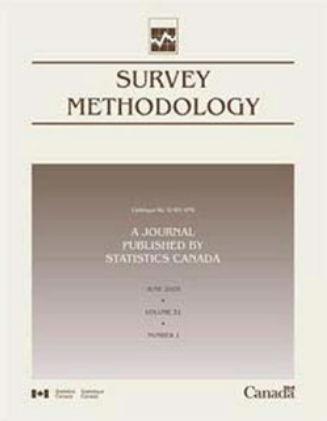
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

From multiple modes for surveys to multiple data sources for estimates



by Constance F. Citro

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 137-161
Statistics Canada, Catalogue No. 12-001-X

From multiple modes for surveys to multiple data sources for estimates

Constance F. Citro[1]

Abstract

Users, funders and providers of official statistics want estimates that are "wider, deeper, quicker, better, cheaper" (channeling Tim Holt, former head of the UK Office for National Statistics), to which I would add "more relevant" and "less burdensome". Since World War II, we have relied heavily on the probability sample survey as the best we could do - and that best being very good - to meet these goals for estimates of household income and unemployment, self-reported health status, time use, crime victimization, business activity, commodity flows, consumer and business expenditures, et al. Faced with secularly declining unit and item response rates and evidence of reporting error, we have responded in many ways, including the use of multiple survey modes, more sophisticated weighting and imputation methods, adaptive design, cognitive testing of survey items, and other means to maintain data quality. For statistics on the business sector, in order to reduce burden and costs, we long ago moved away from relying solely on surveys to produce needed estimates, but, to date, we have not done that for household surveys, at least not in the United States. I argue that we can and must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet user needs from multiple data sources. Such sources include administrative records and, increasingly, transaction and Internet-based data. I provide two examples - household income and plumbing facilities - to illustrate my thesis. I suggest ways to inculcate a culture of official statistics that focuses on the end result of relevant, timely, accurate and cost-effective statistics and treats surveys, along with other data sources, as means to that end.

Key Words: Surveys; Administrative records; Total error; Big data; Income; Housing.

1 Introduction

Tim Holt, former head of the United Kingdom Office for National Statistics and former president of the Royal Statistical Society, once ticked off five formidable challenges for official statistics - namely, to be "wider, deeper, quicker, better, cheaper" (Holt 2007) - to which I would add "less burdensome" and "more relevant". In my view, to respond adequately to one or more, let alone all seven, of these challenges, official statistical offices need to move from the probability sample survey paradigm of the past 75 years to a mixed data source paradigm for the future. Some offices have made that move for most of their statistical programs (see, e.g., Nelson and West (2014) about the extensive use of register-based statistics in Denmark), and almost all offices have made that move for some of their programs, but there are programs not very far along this path. In the case of U.S. household statistical programs, there is a ways to go.

Such a move should not simply elevate another data source as the be all and end all of official statistics in place of the probability sample survey. The 2011 German Republic census - the first census taken in that country since 1983 - provides a useful reminder of the dangers in such an approach. The census results indicated that the administrative records on which Germany based official population statistics for a period of several decades overestimated the population because of failing to adequately record foreign-

1. Constance F. Citro, Director, Committee on National Statistics, U.S. National Academy of Sciences/National Research Council. E-mail: ccitro@nas.edu.

born emigrants (see http://www.nytimes.com/2013/06/01/world/europe/census-shows-new-drop-in-germanys-population.html?_r=0 [November 2014]).

My thesis is that official statistical programs must start with user needs for information for policy development, program evaluation, and understanding societal trends, and work backwards from concepts to appropriate data sources. Such sources may very likely include probability surveys but may also include one or more alternative kinds of data. My thesis is a truism in one sense, but people whose lives are devoted to perfecting a particular tool for data collection may too often see everything as in need of that tool, rather than considering the most cost-effective way to obtain statistics that policy makers, researchers, and other data users want.

I little doubt that Joe Waksberg, whom I was honored to know through his service on a Committee on National Statistics (CNSTAT) Panel on Decennial Census Methodology in the mid-1980s, would approve of my topic. Joe was not only an uncommonly gracious and charming human being, but also a problem-solver and innovator of the first order. Joe stressed "the importance of examining not only what you are asked, but also what you think the analyst has in mind" (Morganstein and Marker 2000). Joe invariably thought outside the box to identify data sources and models that addressed the underlying information need rather than worked from an a priori concept of what tools were appropriate.

In the following text, I briefly review the rise and benefits of probability sampling for official statistics in the United States in Section 2 and the growing threats to the relevance, accuracy, timeliness, cost-effectiveness and public acceptability of survey-based estimates in Section 3. In Section 4 and Section 5, I consider the strengths and weaknesses of administrative records and other non-probability-survey data sources that may be valuable, singly and in combination, for official statistics. In Section 6, I offer examples of ripe opportunities in the United States to transform ongoing household survey programs to use multiple data sources to provide information of greater value. I conclude in Section 7 by enumerating barriers to moving to a multiple data sources paradigm and suggest ways to lower those barriers.

I focus on what I know best - namely, U.S. official statistics and household statistics programs in particular. I hope that readers from other countries, other statistical programs and other agencies will find analogies in their own work. I critique the survey paradigm from a goal of improving official statistics, remaining deeply appreciative of the value of probability surveys, alone and combined with other data sources, and deeply admiring of the important work of statistical agencies in service to the public good (see National Research Council 2013c).

2 The rise of probability sampling in official U.S. statistics

It is not an exaggeration to say that large-scale probability surveys were the 20th-century answer to the need for wider, deeper, quicker, better, cheaper, more relevant and less burdensome official statistics. Such surveys provided information with known precision in contrast to non-probability surveys; and they provided detailed information at greatly reduced cost and increased timeliness compared with censuses. Duncan and Shelton (1978) and Harris-Kojetin (2012) review the rise of probability sampling in U.S. official statistics.

It was not clear at the time when the theory and practice of modern probability sampling was being developed in the 1930s in the United States that probability surveys would gain such widespread

acceptance. The arrival of Jerzy Neyman in the mid-1930s gave a boost to the work of W. Edwards Deming, Calvin Dedrick, Morris Hansen and colleagues at the Census Bureau who were developing the needed theory for sampling of finite populations. Small-scale sample surveys in the 1930s at universities and federal agencies on such topics as consumer purchases, unemployment, urban housing and health provided proofs of concept and practical tips.

Table 2.1
Selected ongoing U.S. statistical agency probability surveys, by year begun

Decade and Year/Type of Survey	Repeated Cross-Sectional Household Survey	Repeated Cross-Sectional Business Establishment Survey	Panel Person Survey
1940	1940 - Current Population Survey (CPS) 1947 - CPS Annual Social and Economic Supplement (CPS/ASEC)	1946 - Monthly Wholesale Trade Survey	
1950	1950 - Consumer Expenditure Survey (CE) 1955 - National Survey of Fishing, Hunting, and Wildlife-Associated Recreation 1957 - National Health Interview Survey (NHIS)	1953 - Advance Monthly Retail Sales Survey 1953 - Business R&D and Innovation Survey (BRDIS) 1959 - Building Permits Survey	
1960	1960 - Decennial Census Long-Form Sample (became American Community Survey in 2005)	1965 - National Hospital Care Survey	1966-1990 - National Longitudinal Survey of Older Men
1970	1972 - National Crime Victimization Survey (NCVS) 1973 - American Housing Survey (AHS); 1973 - National Survey of College Graduates (NSCG) 1979 - Residential Energy Consumption Survey (RECS)	1975 - Farm Costs and Returns Survey and Cropping Practices and Chemical Use Surveys (combined in Agricultural Resource Management Survey in 1996) 1979 - Commercial Buildings Energy Consumption Survey (CBECS)	1972-1986 - National Longitudinal Survey of High School Class of 72 1973-present - Survey of Doctorate Recipients (SDR) 1979-present - National Longitudinal Survey of Youth (NLSY79)
1980	1983 - Survey of Consumer Finances (SCF)	1985 - Manufacturing Energy Consumption Survey (MECS)	1984-present - Survey of Income and Program Participation (SIPP)
1990	1991 - Medicare Current Beneficiary Survey (MCBS)	1996 - Agricultural Resource Management Survey (ARMS)	1997-present - National Longitudinal Survey of Youth (NLSY97)
2000	2005 - American Community Survey (ACS)		2001-2008 - Early Childhood Longitudinal Study (Birth Cohort)

Notes: Current survey name is used; periodicity of interviewing for repeated cross-sectional and panel surveys varies; some repeated cross-sectional surveys have panel component (rotation groups); length of panel surveys (how many years respondents are in sample) varies.
Source: Compiled by author.

The federal government's young statistical Turks still had to surmount hurdles in the bureaucracy up to the White House before they could move sampling into the mainstream of federal statistics. Thus, "old timers" at the Census Bureau were skeptical about the possibility of using survey methods to get information on unemployment and politicians were divided about whether they wanted the estimates (Anderson 1988). In 1937, a major breakthrough occurred when a two percent sample of households on nonbusiness postal routes, designed by Dedrick, Hansen and others, estimated a much higher - and more credible - number of unemployed than a "complete" census of all residential addresses that was conducted on a voluntary basis. Picking up on that effort, from 1940-1942, the Works Progress Administration fielded the sample-based Monthly Report on the Labor Force, the forerunner to the Current Population Survey (CPS). The CPS continues to this day as the source of official monthly estimates of U.S. unemployment conducted by the Census Bureau and published by the Bureau of Labor Statistics (BLS).

Another breakthrough occurred when the Census Bureau, which struggled for decades to respond to demands for added questions on the decennial census without turning the instrument into a nightmare for respondents and interviewers, asked six questions on a five percent sample basis in the 1940 census. The success of sampling led to a decision to administer two-fifths of the questions in the 1950 census to a sample, and subsequent censuses followed suit. Table 2.1 lists selected ongoing U.S. household surveys, business surveys and panel surveys and when they began. The variety of subjects covered and the longevity of these surveys attest to the dominance and value of the sample survey paradigm in U.S. official statistics.

3 Chinks in the armor: Rising threats to the survey paradigm

Probability surveys are indispensable tools for official statistical agencies and others for many kinds of measures - for example, to track such phenomena as public approval of the U.S. president or expressed feelings of well-being. Moreover, probability surveys with a primary purpose to measure constructs, like household income, that could be obtained from other sources, have two major advantages: (1) they can obtain a wide variety of covariates for use in analysis of the primary variable(s) of interest, and (2) they are under the control of the survey designer. Yet threats to the probability survey paradigm are snowballing in ways that bode ill for the future. Manski (2014) goes so far as to accuse statistical agencies of sweeping major problems with their data under the rug and markedly understating the uncertainty in their estimates. He labels survey nonresponse as an example of "permanent uncertainty".

3.1 Characterizing survey quality

A typology of errors and other problems that can compromise the quality of survey estimates is essential for understanding and improving official statistics. A seminal paper in developing data quality frameworks was Brackstone (1999). Most recently, Biemer, Trewin, Bergdahl and Lilli (2014) reviewed the literature on systematic quality frameworks, noting, in particular, the six dimensions proposed by Eurostat (2000): relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability (across time and geography), and coherence (consistent standards). Iwig, Berning, Marck and Prell (2013) reviewed quality frameworks from Eurostat, the Australian Bureau of Statistics, the UK Office for National Statistics, Statistics Canada, and other organizations and developed questions based on six

quality dimensions of their devising - relevance, accessibility, coherence, interpretability, accuracy, and institutional environment - for U.S. statistical agencies to use to assess the utility of administrative records. Daas, Ossen, Tennekes and Nordholt (2012) constructed a framework for evaluating the use of administrative records to produce census data for the Netherlands.

Biemer et al. (2014) went further by using the Eurostat framework (combining comparability and coherence into a single dimension) as the basis for designing, testing and implementing a system of numerical assessments for evaluating and continually improving data product quality at Statistics Sweden. For a full assessment, it would also be necessary to evaluate quality dimensions against cost and respondent burden. Usefully for my purposes, Biemer et al. decomposed the dimension of "accuracy", conceived of as total survey error (or total product error for non-survey-based statistical programs such as national accounts), into sampling error and seven types of nonsampling error: (1) frame error, including undercoverage and overcoverage and missing or erroneous auxiliary variables on the frame; (2) nonresponse error (unit and item); (3) measurement error (overreporting, underreporting, other); (4) data processing error; (5) modeling/estimation error, such as from fitting models for imputation or adjusting data values to conform to benchmarks; (6) revision error (the difference between preliminary and final published estimates); and (7) specification error (the difference between the true, unobservable variable and the observed indicator). For ongoing surveys, I would add *outmoded construct error*, which is related to but different from specification error. For example, the Census Bureau's regular money income concept for official household income and poverty estimates from the CPS Annual Social and Economic Supplement (ASEC) has become progressively outdated due to changing U.S. tax and transfer programs (see, e.g., Czajka and Denmead 2012; National Research Council 1995).

3.2 Four sources of error in U.S. household statistics

3.2.1 Frame deficiencies

Obtaining a comprehensive, accurate frame for surveys can be as difficult as obtaining responses from sample cases drawn from the frame and, in many instances, the difficulties have persisted and even grown over time. Joe Waksberg would resonate to the problem of frame deficiencies: not only did he, with Warren Mitofsky, develop the random digit dialing (RDD) method for generating frames and samples for high-quality residential telephone surveys in the 1970s (see Waksberg 1978; Tourangeau 2004), but he also saw the beginnings of the method's decline in popularity because of such phenomena as cell-phone-only households.

A commonly used frame for U.S. household surveys is the Census Bureau's Master Address File (MAF) developed for the decennial census. The past few censuses have obtained increasingly good net coverage of residential addresses on the MAF, particularly for occupied units (Mule and Konicki 2012). The persistent problem for household surveys is undercoverage of individual members within sampled units. Coverage ratios (i.e., estimates before ratio adjustment to population controls) in the March 2013 CPS, for example, are only 85 percent for the total population, and there are marked differences among men and women, older and younger people, and whites and minorities, with coverage ratios as low as 61 percent for black men and women ages 20-24 (see http://www.census.gov/prod/techdoc/cps/cpsmar13.pdf [November 2014]). No systematic study of the time series of coverage ratios for U.S. household surveys has been conducted, but there is evidence that ratios have been getting worse.

While useful to correct coverage errors for age, gender, race and ethnicity groups, the current household survey ratio adjustments undoubtedly fail to correct for other consequential coverage differences. (The ratio-adjustment controls, in one of the least controversial and most long-standing uses of administrative records in U.S. household surveys, derive from population estimates developed from the previous census updated with administrative records and survey data.) Thus, everything that is known about undercount in the U.S. decennial census indicates that, holding race and ethnicity constant, socioeconomically disadvantaged populations are less well counted than others (see, e.g., National Research Council 2004, App. D). It is unlikely that household surveys perform any better - for example, Czajka, Jacobson and Cody (2004) find that the Survey of Income and Program Participation [SIPP] substantially underrepresents high-income families compared with the Survey of Consumer Finances [SCF], which includes a list sample of high-income households drawn from tax records. Factoring in differential socioeconomic coverage, Shapiro and Kostanich (1988) estimate from simulations that poverty is significantly biased downward for black males in the CPS/ASEC. On the other hand, by comparison with the 2000 census long-form sample, Heckman and LaFontaine (2010) find that survey undercoverage in the 2000 CPS October educational supplement contributes little to underestimates of high school completion rates; other factors are more important.

3.2.2 Unit response in secular decline

A study panel of the (U.S.) National Research Council (2013b) recently completed a comprehensive review of causes and consequences of household survey unit nonresponse, documenting the well-known phenomenon that the public is becoming less available and willing to respond to surveys, even from well-trusted official statistical agencies. In the United States, there was evidence as early as the 1980s that response rates had been declining from almost the beginning of the widespread use of probability sample surveys (see, e.g., Steeh 1981; Bradburn 1992). De Leeuw and De Heer (2002) estimated a secular rate of decline in survey cooperation of 3 percentage points per year from examining ongoing surveys in 16 Western countries from the mid-1980s through the late 1990s. The cooperation rate measures the response of eligible sample cases actually contacted; response rates (there are several accepted variations) have broader denominators, including eligible cases that were not reached (National Research Council 2013c, pp. 9-12). National Research Council (2013b: Tables 1-2, 104) provides initial or screener response rates to a range of U.S. official surveys for 1990/91 (after response rates had already fallen significantly for many surveys) and 2007/2009, which make clear that the problem is not going away.

It was long assumed that lower response rates even with nonresponse weighting adjustments inevitably entailed bias in survey estimates. Recent research (see, e.g., Groves and Peytcheva 2008) finds that the relationship between nonresponse and bias is complex and extraordinary efforts to increase response can inadvertently increase bias by obtaining greater response from only some groups and not others (see, e.g., Fricker and Tourangeau 2010). It would be foolhardy, however, for official statistical agencies to assume that increasing nonresponse has no or little effect on the accuracy of estimates, particularly when unit nonresponse is coupled with item nonresponse. For example, nonrespondents to health surveys are estimated to have poorer health on average than respondents and nonrespondents to volunteering surveys are estimated to be less likely to volunteer than respondents (National Research Council 2013b, pp. 44-45). Moreover, there has been little research on the effects of nonresponse on bivariate or multivariate

associations or on variance, except for the obvious - and not unimportant - effect that unit nonresponse reduces effective sample size.

3.2.3 Item response often low and declining

Neither sample surveys nor censuses can be expected to obtain answers from unit respondents to every item on a questionnaire. U.S. census practice has long been to edit some items for consistency, but until mid-twentieth century, there were no adjustments for item nonresponse - tables included rows labeled “no response” or similar wording. The first use of imputation occurred in 1940 when Deming developed a “cold deck” procedure to impute age by randomly selecting a value for age from an appropriate deck of cards selected according to what other information was known about the person for whom age was missing. Beginning in 1960, with the advent of high-speed computers, “hot deck” imputation methods were used to impute missing values for many census items (Citro 2012). The hot deck procedure uses the latest value for the previously processed person or household stored in a matrix and, consequently, does not have to assume that data are missing completely at random (MCAR), although it does have to assume that data are missing at random (MAR) within the categories defined by variables in the hot deck matrix. Model-based methods of imputation have been developed that do not require such strong assumptions as MAR or MCAR (see National Research Council 2010b), but they are not widely used in U.S. household surveys. Two exceptions are in the Survey of Consumer Finances (SCF) (Kennickell 2011) and the Consumer Expenditure (CE) Interview Survey (Passero 2009).

Whatever the method, imputation has the advantage of creating a full data record for every respondent, which facilitates multivariate analysis and forestalls the likelihood that researchers will use different methods for treating missing data that give different results. Yet imputation may introduce bias into estimates, and the significance of any bias will likely be magnified by the extent of missing data. So it is troubling that nonresponse has been increasing for important items on household surveys, such as income, assets, taxes and consumer expenditures, which require respondents to supply dollar amounts - for example, Czajka (2009:Table A-8) compares item imputation rates for total income and several sources of income for the CPS/ASEC and SIPP for 1993, 1997 and 2002 - a full one-third of income is currently imputed on the CPS/ASEC, up from about one-quarter in 1993 - and SIPP is not much better. Clearly, with such high imputation rates, careful evaluation of the effects of imputation procedures is imperative to carry out. Hoyakem, Bollinger and Ziliak (2014), for example, estimate that the hot deck imputation procedure for earnings in the CPS/ASEC has consistently underestimated poverty by an average of one percentage point, based on evaluating missing earnings in both the CPS/ASEC and Social Security earnings records.

3.2.4 Measurement error problematic and not well studied

Even with complete reporting, or, more commonly, adjustments for unit and item nonresponse, there will still be error in survey estimates from inaccurate reporting by respondents due to guessing at the answer, deliberately failing to provide a correct answer, or not understanding the intent of the question. While acknowledged by statistical agencies, the extent of measurement error is typically less well studied than is sampling error or the extent of missing data. Many measurement error studies compare aggregate estimates from a survey with similar estimates from another survey or an appropriate set of administrative

records, adjusted as far as possible to be comparable. It is not possible to sort out from these studies the part played by measurement error in comparison with other factors, but the results indicate the magnitude of problems. Some studies are able to match individual records and thereby examine components of measurement error.

Significant measurement error is known to affect key socioeconomic estimates produced from U.S. household surveys. Thus, a legion of studies have documented net underestimation of U.S. household income in survey after survey and, even more troubling, a decline in completeness of reporting, even after imputation and weighting. Fixler and Johnson (2012, Table 2), for example, estimated that between 1999 and 2010, mean and median estimates from the CPS/ASEC fell progressively below the National Income and Product Account (NIPA) estimates due to such factors as: (1) underrepresentation of very high-income households in the CPS/ASEC sample; (2) nonreporting and underreporting by those high-income households that are included; and (3) nonreporting and underreporting by middle and lower income households. Studies of individual income sources find even worse error. Meyer and Goerge (2011), for example, by matching Supplemental Nutrition Assistance Program (SNAP) records in two states find that almost 35 percent and 50 percent, respectively, of true recipients do not report receiving benefits in the American Community Survey (ACS) and the CPS/ASEC. Similarly, Meyer, Mok and Sullivan (2009) document large and often increasing discrepancies between survey estimates and appropriately adjusted administrative records estimates of income recipients and total amounts for many sources.

Wealth is notoriously difficult to measure in household surveys, and many do not attempt to do so. Czajka (2009, pp. 143-145) summarizes research on the quality of SIPP estimates of wealth by comparison with the SCF and the Panel Study of Income Dynamics (PSID). Greatly simplifying the findings, SIPP historically has been fairly effective in measuring liabilities, such as mortgage debt, and the value of such assets as owned homes, vehicles, and savings bonds. SIPP has done poorly in measuring the value of assets held mostly by higher income households, such as stocks, mutual funds, and IRA and KEOGH accounts, whereas the PSID has done somewhat better. On net, SIPP significantly underestimates net worth.

A National Research Council (2013a) study of the BLS CE Interview and Diary Surveys found differential quality of reporting of various expenditure types compared with appropriately adjusted personal consumption expenditure (PCE) estimates from the NIPA. Bee, Meyer and Sullivan (2012, Table 2) also find declines in reporting for some expenditures - for example, gasoline reporting in the CE household estimate declined from over 100 percent of the comparable PCE estimate in 1986 to just under 80 percent in 2010, while reporting on furniture and furnishings declined from 77 percent to 44 percent over a comparable period.

4 What can be done?

Survey researchers have not been idle in the face of multiple and increasing threats to the survey paradigm. For at least the last 15 years, they have actively worked on ways to reduce or compensate for coverage error, unit and item nonresponse, measurement error, and, more recently, burden on respondents. Strategies have included: (1) spending more on case completion (although budget constraints limit the viability of this strategy); (2) using paradata and auxiliary information for more effective unit nonresponse

bias identification and adjustment; (3) employing more sophisticated missing data adjustments that do not assume MAR; (4) using adaptive design methods to optimize the cost and quality of response; (5) using multiple frames to reduce coverage error (e.g. cell-phone and land-line frames for telephone surveys); (6) using multiple modes to facilitate more cost-effective response as in the ACS, which recently added an Internet response option to its mail, CATI and CAPI options; (7) reducing burden by optimizing follow-up calls and visits; and (8) describing the needs for the survey data. In the United States, data users are often recruited to make the case to Congress and other stakeholders. For example, the Association of Public Data Users, the Council of Professional Associations on Federal Statistics and the Population Association of America frequently mobilize data users on behalf of statistical agency programs.

My thesis is that these steps, while laudable and necessary, are not sufficient to restore the probability survey-based paradigm for official statistics on households or other types of respondents. I propose, instead, that statistical agencies consistently begin by determining policymakers' and public needs and work backwards to identify appropriate data sources to serve those needs in the most cost-effective and least burdensome manner possible. This multiple sources paradigm should apply to all statistical programs, whether traditionally based on a survey, administrative records, or another source.

Some important statistical programs, such as the NIPAs and the Consumer Price Index (see Horrigan 2013) in the United States and other countries, have for decades used multiple data sources. One reason is that these programs are built around a widely accepted conceptual framework that determines required elements to constitute an acceptable set of estimates. It is not acceptable to omit one or more components of income from the NIPAs simply because data are not available from a single source. Moreover, because key NIPA estimates are periodically revised to add data, improve methodology and refine concepts, there is a built-in positive bias to search for new and improved data sources to fill gaps and improve accuracy. The U.S. economic censuses also use multiple sources, specifically, income tax records for sole proprietors and very small employers together with surveys for larger companies. U.S. household statistics programs, in contrast, have most closely adhered to the probability sample survey paradigm. Moreover, because long intervals typically occur between revisions to household survey concepts and design, the surveys too often fall behind in their ability to serve policymakers and the public, when the use of additional data sources could make possible significant improvements.

5 Which data sources to bolster surveys?

For decades after the introduction of probability sampling in official statistics, the only alternative source was administrative records - from various levels of government, depending on a country's governmental structure (federal, state and local in the United States), and from nongovernmental entities (e. g., employer payroll records or hospital admission records). And a number of national statistical agencies around the world began to incorporate administrative records into their programs - from using them in an ancillary way to moving census and survey programs lock, stock and barrel to an administrative records-based paradigm.

Technological innovations in the 1970s and 1980s led to some additional data sources - such as records of expenditures at checkouts (made possible by the development of bar codes and scanners) and aerial and satellite images for categorizing land use - becoming at least potentially available for official statistics. But the landscape of data sources was still relatively contained. Beginning in the 1990s, the advent of the

Internet and high-speed distributed computing technology unleashed a mind-boggling array of new data sources, such as data from traffic camera feeds, tracking of cell phone locations, search terms used on the Web and postings on social media sites. The challenge for statistical agencies is to classify and evaluate all of these data sources in ways that help agencies determine their usefulness.

5.1 Is "Big Data" a useful concept?

Many new types of data that have become available in the past 15 or so years are often very large in size, leading to the use of the term "big data". I argue that this buzz phrase does little, if anything, to assist statistical agencies to determine appropriate combinations of data for their programs. In computer science, "big data is high volume, high velocity and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization" (Laney 2001). These properties are not inherent in any particular type of data or in any particular platform, such as the Internet. Instead, what qualifies as "big data" is a changing target, as advances are made in high-speed computing and data analysis techniques. In today's computing environment, census, survey, and administrative records data rarely qualify as "big", although they may have done so in an earlier era. People today tend to classify as "big" the data streams from cameras, sensors, and largely free-form interactions with the Internet, such as social media postings. In the future, many of these kinds of data may no longer fit under this rubric. In regard to the Internet, moreover, it not only generates a great deal of today's "big data", but also provides ordinary-size data in a more accessible way - for example, access to public opinion polls or to local property records.

I would argue that statistical agencies will most often want to be and should be "close followers" rather than leaders in using big data. It seems to me most appropriate for academia and the private sector to be out front in tackling the uses of data that are so voluminous and of such high velocity and variety that they require big leaps forward to develop new forms of processing and analysis. Statistical agencies should be alert to developments in the field of big data that promise benefits for their programs down the road and may be well advised to support research in this area to help ensure that applications that are relevant to their programs emerge. Principally, however, I believe that statistical agency resources are best used primarily for working with data sources that offer more immediately useful benefits.

Groves (2011) has attempted to move toward a more relevant classification for statistical agencies than that between "big data" and all other data, by distinguishing between what he terms "designed data" that are "produced to discover the unmeasured" and "organic data" that are "produced auxiliary to processes, to record the process". Keller, Koonin and Shipp (2012) list examples of data sources under Groves' two headings. Their list of designed data includes: administrative data (e.g., tax records); federal surveys; censuses of population; and "other data collected to answer specific policy questions". Their list of organic data includes: location data (cell phone "externals", E-ZPass transponders, surveillance cameras); political preferences (voter registration records, voting in primaries, political party contributions); commercial information (credit card transactions, property sales, online searches, radio-frequency identification); health information (electronic medical records, hospital admittances, devices to monitor vital signs, pharmacy sales); and other organic data (optical, infrared and spectral imagery, meteorological measurements, seismic and acoustic measurements, biological and chemical ionizing radiation). Not mentioned under either category are such data as Facebook or Twitter postings, although they might fall under the broad rubric of "online searches".

Whether the two-part classification in Keller et al. (2012) is all that more useful than "big data" for statistical agency purposes is a question. For example, classifying voter registration records or electronic health records as organic data and not as designed administrative data seems to miss ways in which they differ from such sources as online searches and ways in which they are similar to federal and state government administrative records. Moreover, even organic data are "designed", if only minimally, in the sense that the provider has specified some parameters, such as 140 characters for a Twitter post or a particular angle of vision for a traffic camera. Nonetheless, the designed versus organic distinction does point to a useful dimension, which is the degree to which statistical agencies have ready access to, control changes to, and are able readily to understand the properties of a data source.

5.2 Dimensions of data sources: Illustrations for four major categories

Coming up with satisfactory nomenclature and evaluation criteria that can help statistical agencies assess the potential usefulness of alternative data sources for their programs, with the goal of becoming as familiar with the error properties of alternative sources as they are with total error for surveys, is not going to happen without considerable effort by statistical agencies around the world (Iwig et al. 2013 and Daas et al. 2012 are examples of such efforts). I do not pretend that I can come close to that goal in this paper. My goal is more modest - namely, to provide some illustrations so that those who are wedded to a probability survey paradigm (or an administrative records paradigm) can see that the task of understanding alternative data sources is both feasible and desirable. I provide illustrations for four data sources ranging from traditional to cutting-edge:

(1) Surveys and censuses, or a collection of data obtained from responses of individuals, who are queried on one or more topics as designed by the data collector (statistical agency, other government agency, and academic or commercial survey organization) according to principles of survey research with the goal of producing generalizable information for a defined population.

(2) Administrative records or a collection of data obtained from forms designed by an administrative body according to law, regulation, or policy for operating a program, such as paying benefits to eligible recipients or meeting payroll. Administrative records are usually ongoing and may be operated by government agencies, or non-governmental organizations.

(3) Commercial transaction records, or a collection of data obtained from electronic capture of purchases (e.g., groceries, real estate) initiated by a buyer but in a form determined by a seller (e.g., bar-coded product information and prices recorded by check-out scanners or records of product and price information for Web sales, such as through Amazon).

(4) Interactions of individuals with the WorldWide Web by using commercially provided tools, such as a Web browser or social media site. This category covers a wide and ever-changing array of potential data sources for which there are no straightforward classifications. One defining characteristic is that individuals providing information, such as a Twitter post, act as autonomous agents: they are not asked to respond to a questionnaire or required to supply administrative information but, instead, are choosing to initiate an interaction.

I first rank each source on the following two dimensions, which relate to the framework in Biemer et al. (2014). The rank I assign assumes there have been as yet no proactive steps by a statistical agency to

boost the ranking (e.g., by embedding staff in an administrative agency to become deeply familiar with the agency's records). The two dimensions are:

(1) Degree of accessibility to and control by national statistical agency: high (statistical agency designs the data source and controls changes to it); medium (statistical agency has authority to use the data source and influence on changes to it); low (statistical agency must arrange to obtain the data source on the terms of the provider and has little or no influence on changes to it). Gradations can be added to each of these categories depending, for example, on how strong an agency's authority is to acquire a set of administrative records.

(2) Degree to which components of error can be identified and measured: high, as in designed surveys and censuses; medium, as in public and private sector administrative records; and low, as in streams of data from autonomous choices of individuals.

I further identify aspects of data quality for each source, following Biemer et al. (2014). I also indicate variations for most of the dimensions depending on the provider, such as national statistical agency, other unit of national government, other level of government, academic institution, or commercial entity. Table 5.1 provides all of this information as best I can.

An ideal source for statistical agency use, other things equal, is one that is provided, designed, and controlled by the agency, and for which errors can be identified and measured and are generally under control, such as a high-quality probability survey mounted by the agency. At the other extreme is a data source that is controlled by one or more private companies (e.g., scanner data) or, perhaps, by hundreds or thousands of local governments (e.g., traffic cameras), where the data result from autonomous choices or uncontrolled movements, and where it is difficult to conceptualize, much less measure, errors in the data source. Yet when considering a statistical agency's responsibility to provide relevant, timely, accurate statistics for policymakers and the public for which costs and respondent burden are minimized, there may well be non-survey data sources that warrant the effort to make them usable for statistical purposes. I argue that the threats to the survey paradigm reviewed above make it imperative to consider alternative data sources because surveys are no longer always and everywhere demonstrably the superior choice to other sources - they are not always "high" on the dimensions in Table 5.1.

I further argue that government administrative records, which, as Table 5.1 indicates, more often have desirable properties for official statistics compared with other non-survey data sources, should be a prime candidate for statistical agencies to incorporate as extensively as possible into their survey programs if they have not already done so. Administrative records are generated according to rules - rules about the eligible population, who must file what information, what action by the pertinent administrative body is taken on the basis of the information (e.g., tax refund, benefit payment), and so on. This fact should make it possible, with requisite effort, for a statistical agency to become as familiar with administrative records error structures as they are with total survey error. Couper (2013) provides a useful discussion somewhat like mine. He pokes holes in the ability of organic data sources to be as useful as they are often touted to be, much less to be suitable to replace probability surveys, but he warns survey researchers that they ignore organic data sources at their peril. Ironically, his conclusion to make some use of organic sources is strengthened because of his error in classifying administrative records as organic data. They are properly classified as designed data, even though not designed by a statistical agency.

Table 5.1
Ranking (HIGH, MEDIUM, LOW, VERY LOW, or VARIES) of four data sources on dimensions for use in official statistics

Dimension/ Data Source	Census/Probability Survey (e.g., CPS/ASEC, ACS, NHIS - see Table 2.1)	Administrative Records (e.g., income taxes, Social Security, unemployment, payroll)	Commercial Transaction Records (e.g., scanner data, credit card data)	Individual Interactions with the Internet (e.g., Twitter postings; Google search term volumes)
Degree of Control by/ Accessibility to Statistical Agency	HIGH (survey conducted for statistical agency); MEDIUM to LOW (survey conducted for private organization)	HIGH to MEDIUM (national agency records); MEDIUM to LOW (state or local records); MEDIUM to LOW (commercial records)	MEDIUM to LOW	VERY LOW
Degree of Ability of Statistical Agency to Identify/Assess Properties/ Errors	HIGH (survey conducted for statistical agency); VARIES (survey conducted for private org., depends on documentation and transparency)	HIGH to MEDIUM (national agency records); MEDIUM to LOW (state or local records); MEDIUM to LOW (commercial records)	MEDIUM (to the extent that records follow accepted standards, e.g., for bar coding and pricing information)	VERY LOW
Data Quality Attributes (Biemer et al. 2014)				
Relevance for Policy and Public - Concepts and Measures	HIGH for survey conducted for statistical agency, assuming well designed and up to date in concepts and measures; VARIES for surveys for private organizations	VARIES across and within records systems (e.g., records of benefit payment may be highly relevant, while family composition information may use a different concept)	VARIES	VARIES, but VERY LOW at the present state of the art of acquiring, evaluating, and analyzing these kinds of data
Relevance - Useful Covariates	HIGH for most surveys	VARIES, but rarely as high as for most surveys	VARIES, but rarely as high as for most surveys	VARIES, but typically LOW
Frequency of Data Collection	Weekly to every few years (every decade for the U.S. population census); Some private surveys, such as election polls, may run daily	Generally records are updated frequently (e.g., daily) and continually	Generally records are updated frequently (e.g., at moment of transaction or daily) and continually	Interactions are captured instantaneously
Timeliness of Release	VARIES, depending on effort of statistical agency or private organization, but some lag from the reference period for responses is inevitable	VARIES, but some lag from the reference date to when records are acquired by statistical agency likely	VARIES, but likely to be long lags in acquiring proprietary data by statistical agency	VARIES, but likely to be long lags (although MIT Billion Prices Project has worked out very timely access for prices on the Internet; see bpp.mit.edu)
Comparability and Coherence	HIGH across time and geography within survey (except when deliberately changed or if societal change that affects measurement is not taken into account); VARIES among surveys	HIGH within records system (changes to government records generally heralded by legal/ regulation/policy change, changes to commercial records likely opaque); VARIES among records systems	HIGH within records system (changes generally opaque to statistical agency); VARIES among records systems	VERY LOW, in that vendors (e.g.,Twitter) may add/subtract features or drop an entire product; Changes generally opaque to statistical agency; Initiators of interactions may have very different frames of reference

	Accuracy (Components of Error)*			
Dimension/ Data Source	**Census/Probability Survey (e.g., CPS/ASEC, ACS, NHIS - see Table 2.1)**	**Administrative Records (e.g., income taxes, Social Security, unemployment, payroll)**	**Commercial Transaction Records (e.g., scanner data, credit card data)**	**Individual Interactions with the Internet (e.g., Twitter postings; Google search term volumes)**
Frame Error	VARIES, can be significant undercoverage and overcoverage	Frame is usually well defined by law, regulation, or policy; Problem for statistical agency use is that frame may not be comprehensive	Frame is ill-defined for statistical agency purposes, in that represents whoever had a purchase scanned by a specified vendor or used a specific credit card for a purchase during a specified time; Poses significant challenge to statistical agency to determine appropriate use	Frame is ill-defined for statistical agency purposes, in that represents whoever, decided to, for example, set up Twitter account or conduct Google search during a specified time; Poses significant challenge to statistical agency to determine appropriate use
Nonresponse (unit and item)	VARIES, can be significant	VARIES (e.g., Social Security records likely to include almost all eligible people, but income tax records likely to reflect evasion, in terms of failure to file a return or concealing some income)	NOT APPLICABLE, in that "respondents" are self selected; Statistical agency challenge is to determine appropriate use that does not need to assume a probability mechanism	NOT APPLICABLE, in that "respondents" are self selected; Statistical agency challenge is to determine appropriate use that does not need to assume a probability mechanism
Measurement Error	VARIES within surveys by item and among surveys for comparable items; Often not well assessed, even for statistical agency surveys	VARIES among record systems and within record systems by item depending on centrality of the item to program operation (e.g., benefit payment item likely more accurate than items obtained from beneficiaries, such as employment status)	NOT APPLICABLE to data source as such, although any characteristics added by the vendor from another source may/may not be valid; Statistical agency challenge is to not introduce measurement error by inappropriate use of the data	NOT APPLICABLE to data source as such, although any characteristics added by the vendor from another source may/may not be valid; Statistical agency challenge is to not introduce measurement error by inappropriate use of the data
Data Processing Error	VARIES (e.g., may be data capture or recoding errors), but is usually under good statistical control, although harder to assess for private organization surveys	VARIES (e.g., may be keying or coding errors), likely to be under better control for key variables (e.g., benefit payments) than for other variables, but hard for statistical agency to assess	VARIES (e.g., may be errors in assigning bar codes or prices), likely to be under good control, but hard for statistical agency to assess	NOT APPLICABLE, in that error is not defined, although there may be occasional problems of the sort that, say, a day's worth of Twitter posts is overwritten and lost
Modeling/ Estimation Error	Bias from such processes as weighting and imputation VARIES; Often intense effort by statistical agency to design well initially but not to revisit to ascertain continued validity of procedures	NOT APPLICABLE (usually), in that records are "raw" data, except perhaps for some recoded variables, but bias may be introduced by statistical agency reprocessing	NOT APPLICABLE (usually), in that records are "raw" data, except perhaps for some recoded or summarized variables, but bias may be introduced by statistical agency reprocessing	NOT APPLICABLE (usually), in that records are "raw" data, but statistical agency reprocessing may introduce significant bias (e.g., by using the word "fired" as always indicating unemployment in analyzing Twitter posts)

Accuracy (Components of Error)* (CON'T)				
Dimension/ Data Source	**Census/Probability Survey (e.g., CPS/ASEC, ACS, NHIS - see Table 2.1)**	**Administrative Records (e.g., income taxes, Social Security, unemployment, payroll)**	**Commercial Transaction Records (e.g., scanner data, credit card data)**	**Individual Interactions with the Internet (e.g., Twitter postings; Google search term volumes)**
Specification Error	VARIES (e.g., self-reported health status may validly indicate respondent's perception but not necessarily diagnosed physical or mental health); May change over time (e.g., as word usage changes among the public)	VARIES; can be significant when administrative records concept differs from what statistical agency needs (e.g., rules for reporting earnings on tax forms may leave out such components as cafeteria benefits)	VARIES; can be low or high depending on how well the data correspond to statistical agency needs	VARIES, but likely significant at the present state of the art of acquiring, evaluating, and analyzing these kinds of data that arise from relatively free-form choices of autonomous individuals
Burden*	VARIES, can be high	NO ADDITIONAL BURDEN from statistical agency on relevant population (e.g., beneficiaries), but burden on administrative agency	NO ADDITIONAL BURDEN from statistical agency on relevant population (e.g., shoppers), but burden on vendor	NO ADDITIONAL BURDEN from statistical agency on relevant population (e.g., Twitter posters), but burden on vendor
Cost*	VARIES, can be high; Statistical agency bears full costs of design, collection, processing, estimation	VARIES, but could be lower than comparable survey because administrative agency bears data collection costs, but statistical agency likely incurs costs of special processing/ handling	VARIES as for administrative records, but vendor likely to want payment; Statistical agency likely incurs costs of special processing/ handling/ analyzing	VARIES as for administrative records, but vendor likely to want payment; Additional statistical agency costs for processing/analyzing unstructured data may be high

*Direction of scale changes; that is "high" is undesirable and "low" is desirable.
Note: Excludes revision error from the Biemer et al. (2014) classification.
Source: Author's rough assessment.

5.3 Uses of administrative records for household survey-based programs

Household survey respondents have demonstrated time and time again that their responses to many important questions on income, wealth, expenditures, and other topics are not very accurate. Use of administrative records has the potential in many instances to remedy this situation. An alternative strategy of many U.S. household survey programs has been to encourage the respondents themselves to consult their own records, such as tax returns, when answering questions on income and similar topics. Certainly, answers are likely to be more accurate when records are consulted, as Johnson and Moore (no date) find in a comparison of income tax records with SCF responses for the 2000 tax year. However, the strategy itself appears to be largely an exercise in futility. The same study of the SCF by Johnson and Moore reports that only ten percent of households with an adjusted gross income of less than $50,000 consulted records and that only 22 percent of higher income households did so. See National Research Council (2013a, pp. 89-91) and Moore, Marquis and Bogen (1996) for similar findings about the difficulties of getting respondents to consult records.

Turning to strategies for statistical agencies to work with administrative data directly, I identify eight ways in which administrative records can contribute to household survey data quality: (1) assist in evaluation of survey data quality, by comparison with aggregate estimates, appropriately adjusted for differences in population universes and concepts, and by exact matches of survey and administrative records; (2) provide control totals for adjusting survey weights for coverage errors; (3) provide supplemental sampling frames for use in a multiple frame design; (4) provide additional information to append to matched survey records to enhance the relevance and usefulness of the data; (5) provide covariates for model-based estimates for smaller geographic areas than the survey can support directly; (6) improve models for imputations for missing data in survey records; (7) replace "no" for survey respondents who should have reported an item, replace "yes" for survey respondents who should not have reported an item, and replace reported values for survey respondents who misreport an item; and (8) replace survey questions and use administrative records values directly. In a longer unpublished version of this article, I provide some current and potential examples of each type of use and identify benefits, confidentiality and public perception concerns, and limitations and feasibility issues for each use generically and specifically for U.S. household surveys on such topics as income, assets and expenditures. My bottom line is that the benefits should outweigh the drawbacks, given a sustained program to integrate administrative records systems with statistical programs.

5.4 Potential uses of non-traditional data sources

Having previously indicated that data from sources other than surveys and administrative records are problematic in a number of ways for official statistics, I would be remiss not to discuss briefly why such data appear to be so attractive. Private companies have very different loss functions from statistical agencies - they are seeking an edge over competitors. Data that are more timely and that identify ways to increase sales and profits are likely useful to a private company, even if they do not cover a population completely or have other drawbacks for official statistics. From this perspective, the kinds of experiments that a Google does, using its own "big data", on ways to increase ad views are good investments (see, e.g., McGuire, Manyika and Chui 2012). Similarly, program agencies at all levels of government, often working with academic centers, are putting together and analyzing their own and other data in innovative ways to identify patterns, "hot spots", and the like, not only for improving their programs and planning new services, but also for prioritizing resources and improving response in real time (see, e.g., the Center for Urban Science and Progress at New York University (http://cusp.nyu.edu/); and the Urban Center for Computation and Data at the University of Chicago (https://urbanccd.org))

Statistical agencies need, above all, sources of data that cover a known population with error properties that are reasonably well understood and that are not likely to change under their feet - characteristics that are not inherent in such data sources as autonomous interactions with websites on the Internet. There are, however, at least two ways in which household survey-based statistical agency programs could obtain an "edge" from non-traditional sources: one is to improve timeliness for preliminary estimates of key statistics; and the other is to provide leading indicators of social change (e.g., the emergence of new occupations and fields of training) that alert statistical agencies to needed changes in their concepts and measures.

6 From data needs to data sources: Two U.S. examples

For concreteness, I offer two U.S. examples - household income and housing unit characteristics - where I believe it is possible and incumbent on statistical agencies to turn survey programs into multiple sources programs to best meet user needs. The U.S. Office of Management and Budget (2014) has taken a positive step in this direction in a recent memorandum asserting that statistical uses of federal agency administrative records are a positive good and outlining step to institutionalize their use.

6.1 Household income

Official statistics on the distribution of household income are among the most important indicators of economic well-being that are regularly produced by national statistical offices, and they are even more important in light of today's debates about rising inequality and related topics. Yet it is abundantly clear that the quality of household income measures obtained from responses to U.S. surveys is significantly impaired by coverage error, unit nonresponse, item nonresponse and misreporting. Moreover, the concept of regular money income for U.S. surveys is out of date with respect to the complex and continually evolving ways in which households obtain resources for everyday consumption and savings. It seems imperative for the U.S. statistical system to improve its flagship income estimates from CPS/ASEC, SIPP, and, to the extent feasible, the ACS by moving from relying largely on survey responses to an approach that integrates survey and administrative records data. The Census Bureau is implementing new and modified questions to better measure retirement income and other sources in the CPS/ASEC, consequent to a major review of income measurement in that survey by Czajka and Denmead (2012) and a report on cognitive testing of changes to the ASEC questionnaire (Hicks and Kerwin 2011). The Census Bureau also recently implemented a major redesign of SIPP, using event history calendar methods and annual interviews in place of interviews every four months to reduce burden and costs, with effects on quality to be evaluated (see https://www.census.gov/programs-surveys/sipp/about/re-engineered-sipp.html [November 2014]). There is a process in place for review of questions on the ACS, although as yet questions on income have not been tackled. The flagship surveys would be markedly improved if, in addition to continued standard questionnaire research to identify ways to reduce burden, clarify question meaning, and facilitate response to the income questions to the extent possible, the following four steps were taken:

(1) The U.S. Census Bureau and Bureau of Economic Analysis (BEA) were to agree on - and periodically revisit and update as appropriate - a contemporary concept of regular household income on which to base estimates from the CPS/ASEC, SIPP and ACS, and the personal income series in the NIPAs, which are developed largely from administrative records. The surveys and NIPAs currently have conceptual differences, such as in the treatment of retirement benefits, which should be reconciled. Using an integrated concept of household income would make both the personal income accounts and household surveys more useful for analyzing trends from macro and micro perspectives.

(2) The Census Bureau was to conduct research on the likely benefits from implementing socioeconomic survey weight adjustments in addition to demographic weight adjustments.

Assuming a benefit, the Census Bureau would next identify appropriate sources, which could be income tax records or the SCF, to adjust weights in the CPS, SIPP and the ACS to capture coverage differences by broad socioeconomic class.

(3) The Census Bureau was to move strategically, source by source, to improve imputations of income receipt and amounts in the CPS/ASEC and SIPP by using administrative records values. The Census Bureau already has access to many records and is working to obtain additional records (e.g., SNAP records from states) as part of 2020 census planning.

(4) The Census Bureau was to move - carefully, in consideration of the added hurdles for use of administrative records in the United States - toward the Statistics Canada model, whereby respondents can skip entire blocks of income questions by permitting access to their income tax and other administrative records (see http://www.statcan.gc.ca/eng/survey/household/5200 [November 2014]).

I do not mean to underestimate the difficulties of the steps outlined above for U.S. income statistics. These difficulties, in no particular order, include: (1) legal and bureaucratic impediments to obtaining ready access to administrative records, which are orders of magnitude greater for records held by state agencies because of differences in state laws, policies and data standards and systems; (2) respondent consent considerations, particularly if values from records are substituted for questions; (3) perceptions of "big brother" and threats to privacy, which may limit the accessibility of microdata for research and policy analysis; (4) lack of resources for statistical agencies to initiate such activities as redesign of imputation systems; (5) adverse effects on timeliness to the extent that records lag in their availability to statistical agencies, which could be addressed by issuing preliminary estimates followed by final estimates when sufficient administrative data become available; (6) insufficient knowledge of the error structures of records, which could lead to nasty surprises; (7) differences in concepts between records and survey measures that are not readily addressed (e.g., earnings reported to the IRS are not gross earnings but earnings subject to tax); (8) additional burdens on already-stretched-thin statistical agency headquarters staff; (9) the need to rewrite processing systems to link multiple streams of data and conduct all needed matching, reconciling and estimation on a timely basis; (10) the distrust of many U.S. microdata users, who seem to prefer a single-source data set, such as a survey, regardless of inaccuracies in the data, to a multiple-source data set, which may include model-based values for some variables; and (11) the hesitation of statistical agency staff, who often seem to believe that it is improper to use, say, administrative records to impute income receipt to a respondent who did not indicate receipt or to use administrative records to substitute for some questions or improve some imputations, unless this can be done for all items. In planning for the 2020 census, the Census Bureau is considering limited use of administrative records for nonresponse follow-up, which could be a model for selected use of records in household surveys. Although formidable, none of these difficulties are insurmountable. A well-articulated, staged, strategic plan for taking a multiple sources approach could empower statistical agencies to work toward quality gains for income estimates and achieve at the same time a reduction in respondent burden and potentially a reduction in costs for key survey programs.

6.2 Housing characteristics, including plumbing facilities

Originating in the New Deal's concern with poor housing quality for much of the nation, the 1940 U.S. decennial census included a few questions on the characteristics of housing units. That concern was well founded - the 1940 census found, for example, that 45 percent of housing units lacked complete plumbing facilities (hot and cold piped water, flush toilet, shower or bathtub). See https://www.census.gov/hhes/www/housing/census/historic/plumbing.html [November 2014]. Housing questions grew in number and were included on censuses through 2000. When the American Community Survey came on-line, it included the housing questions previously on the long-form sample. A much smaller biannual American Housing Survey (AHS) collects an even wider range of information about housing and neighborhoods.

The major reason to investigate ways to move the ACS housing questions from a survey-based program to a survey-plus-alternative-data-sources-based program is respondent burden, both actual and perceived, which in the current political climate in the United States threatens the viability of the ACS. Because the ACS is in the field with a large sample of about 280,000 households every month, instead of once every ten years as for the census long-form sample that it replaced, the survey generates a small but continuous stream of complaints to members of Congress, which have led to congressional hearings. The Census Bureau has identified four items on the ACS that give rise to the most complaints - income, disability, time of leaving for work and plumbing facilities (see http://www.census.gov/acs/www/Downloads/operations_admin/2014_content_review/ACSContentReviewSummit.pdf[November 2014]). The questions on plumbing facilities in the census long-form sample were also regularly the butt of jokes and complaints. In fact, people answer these questions quite completely (see http://www.census.gov/acs/www/methodology/item_allocation_rates_data/ [October 2014]), yet the questions continue to be resented and sometimes not well understood (see Woodward, Wilson and Chestnut 2007). Moreover, an examination of the full ACS questionnaire suggests that many households experience a substantial burden from the total set of about 30 housing questions, particularly homeowners with a mortgage.

The Census Bureau responded to the concerns about ACS burden by cutting back the number of follow-up calls and visits (see Zelenak and Davis 2013), establishing a "respondent advocate", and giving the public information about the rationale for the questions. Yet the U.S. House of Representatives passed an appropriations bill May 30, 2014, that, if enacted, would turn the ACS into a voluntary instead of mandatory survey. While good quality data could likely be collected given enough follow-up, the costs of the ACS would increase substantially (see Griffin 2011). The Census Bureau recently asked federal agencies about legislative or regulatory justification for each and every question, with the real possibility that some questions will be dropped (see http://www.census.gov/acs/www/about_the_survey/acs_content_review/[November 2014]). Plumbing facilities might seem to be a good candidate for deletion from the ACS, given that only 0.4 percent of U.S. housing units lacked complete plumbing facilities in 2012 (From http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_1YR_DP04&prodType=table [November 2014]). However, that small percentage is concentrated in particular areas, such as Native American reservations and rural areas. Moreover, deleting any question on the ACS seems a drastic step to take without first exploring whether alternative sources might provide the data.

In fact, there are housing items on the ACS questionnaire that could very likely be obtained from a variety of other sources, attached to the Census Bureau's Master Address File (MAF), and be available for inclusion in the ACS and other surveys that use the MAF as a sampling frame. Alternative sources include local government administrative records of taxes assessed, year built, and other characteristics of properties, which are increasingly being compiled by commercial vendors, thereby reducing the need to interact with the thousands of individual governments in the United States. They also include sources like Google Street View for exterior property characteristics, realtor websites for housing value and interior characteristics (e.g., number of rooms), smart meters for utility costs (in use in some areas and likely to spread in the future), and mortgage databases held by federal agencies and commercial vendors. Housing characteristics that rarely change can also be drawn from the previous census long-form samples. Plumbing facilities is a prime example - once a house is plumbed, it is almost never unplumbed (even though at times the plumbing may not be functional).

These alternative sources will vary in how easily they are acquired and evaluated, the actual or perceived lack of threat to privacy and confidentiality they pose, and the extent to which they cover all or most of the country. Development of an augmented Master Address and Housing Unit File (MAHUF) that can serve the ACS and other Census Bureau statistical programs will take time, and, for some items (e.g., plumbing facilities), it may be necessary to use a separate (longer) version of the questionnaire in selected geographic areas that appends the relevant items. All of this will be messy, but the long-term potential payoffs are substantial. To move toward an augmented MAHUF, the Census Bureau can benefit from work of the Office of Policy Development & Research in the U.S. Department of Housing and Urban Development to streamline the lengthy AHS questionnaire by using other sources of data for many housing and neighborhood characteristics in place of survey questions; see http://www.huduser.org/portal/datasets/ahs.html#planning [November 2014].

7 Challenges and strategies for effecting paradigm change

I have argued for a paradigm change in which statistical agencies design and update their flagship programs by determining the best combination of data sources and methods to serve user needs in a topic area of ongoing importance. I use U.S. household surveys as an example, where the evidence is strong that relying on survey responses alone will not suffice to serve critical needs for high-quality information on income, expenditures, and related subjects. I expect it is also true that the use of administrative records alone, as in some countries with detailed population registers, may not provide sufficiently complete and high-quality information in the absence of regular efforts to review the quality of the register data and augment and correct them with information from other sources, such as surveys. As a case in point, Axelson, Homberg, Jansson, Werner and Westling (2012) describe the utility of surveys for evaluating the quality of housing and household data from a new dwelling register that was constructed for the 2011 census in Sweden.

I close by listing factors that make paradigm change difficult, countered by ways to effect the change I recommend and ingrain it in statistical agency culture. The U.S. and other statistical systems have admirable records of innovation in many aspects of their programs, but changing paradigms is always difficult, as was evident in the battle to introduce probability sampling to official U.S. statistics in the

1930s. It is particularly hard to rethink long-lived, ongoing, statistical programs with which both the producer agency and the user base are comfortable.

Factors that can impede change include: (1) inertia, particularly when a program was originally innovative and very well designed, so it can coast on its earlier success; (2) becoming out of touch with stakeholders' changing needs, which can be exacerbated when an agency views itself as the only source of needed data and not in competition; (3) fear of undercutting existing programs combined with fear of "not-invented here"; (4) inadequate ongoing evaluation of data quality in all of its dimensions; and (5) constrained staff and budget resources, coupled with an understandable reluctance of agency staff or their established user base to cut back on one or another long-standing statistical series in order to make important advances in other series.

Yet there are many outstanding examples of important innovation in U.S. and other nation's statistical agencies, so clearly there are ways to overcome the constraints listed above to effect paradigm change. The essential ingredient for paradigm change, I believe, is leadership buy-in and continued support at the top of a statistical agency, proactively deployed to garner buy-in at all levels of the agency. For an outstanding example of such leadership, see the discussion in National Research Council (2010a) of the role of Morris Hansen and his colleagues in reengineering what had been an enumerator-based census into a mailout/mailback census. The reengineering effort was initiated and sustained on the basis of evidence of substantial interviewer bias and variance for important data items. There was also concern that it could become more difficult to recruit enumerators as women moved into the work force.

Specific steps for agency leadership to get behind for the specific purpose of inculcating the use of multiple data sources for ongoing official statistical programs include (see Prell, Bradsher-Fredrick, Comisarow, Cornman, Cox, Denbaly, Martinez, Sabol and Vile (2009), who conducted case studies of successful statistical uses of administrative records in the United States, for similar conclusions): (1) setting clear expectations and goals for staff, such as the expectation that statistical programs will, as a matter of course, combine such sources as surveys and administrative records in the interests of relevant, accurate and timely data produced cost-effectively and with minimal respondent burden; (2) according a prominent role to subject-matter specialists - to interface with outside users and inside data producers; (3) staffing operational programs with expertise in all relevant data sources, which includes putting specialists in survey design and specialists in administrative records or other data sources on an equal footing; (4) providing for rotation of assignments, including internal rotations, rotations among statistical agencies, rotations with data user organizations and rotations with sources of alternative data sources; (5) carving out resources for continued evaluation; and (6) treating organizations with alternative data sources that play important roles in statistical programs as partners. On this last point, see, e.g., Hendriks (2012, p. 1473), who, in discussing the experiences of Statistics Norway with their first register-based census in 2011, stresses that "The three C's of register based statistics (in order to achieve data quality) are Co-operation, Communication and Coordination."

Statistical agencies have shown the ability to make far-reaching changes in response to threats to established ways of doing business. The second half of the 20th century gave us the probability survey paradigm in response to the increasing costs and burden of conducting full enumerations and the flaws of non-probability designs. The 21st century can surely give us the paradigm of using the best source(s), including surveys, administrative records and other sources, to respond to policy and public needs for relevant, accurate, timely and cost-effective official statistics.

Acknowledgements

This paper is based on the author's years of experience at the Committee on National Statistics, but the views expressed are her own and should not be assumed to represent the views of CNSTAT or the National Academy of Sciences. The author thanks John Czajka, David Johnson and Rochelle Martinez for helpful comments on an earlier draft. A longer version of this paper is available from the author on request.

References

Anderson, M.J. (1988). *The American Census: A Social History*. New Haven. CT: Yale University Press.

Axelson, M., Homberg, A, Jansson, I., Werner, P. and Westling, S. (2012). Doing a register-based census for the first time: The Swedish experience. *Paper presented at the Joint Statistical Meetings*, San Diego, CA (August). Statistics Sweden, Stockholm.

Bee, A., Meyer, B.D. and Sullivan, J.X. (2012). The validity of consumption data: Are the consumer expenditure interview and diary surveys informative? *NBER Working Paper No. 18308*. Cambridge, MA: National Bureau of Economic Research.

Biemer, P., Trewin, D., Bergdahl, H. and Lilli, J. (2014). A system for managing the quality of official statistics, with discussion. *Journal of Official Statistics*, 30(3, September), 381-442.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25(2), 139-149.

Bradburn, N.H. (1992). A response to the nonresponse problem. 1992 AAPOR Presidential Address. *Public Opinion Quarterly*, 56(3), 391-397.

Citro, C.F. (2012). *Editing, Imputation and Weighting*. Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey, Second Edition, M. J. Anderson, C.F. Citro and J.J. Salvo, eds, 201-204. Washington, DC: CQ Press.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Keynote presentation at the *5th European Survey Research Association Conference*. Ljubliana, Slovenia. http://www.europeansurveyresearch.org/sites/default/files/files/Couper%20keynote.pdf [July 2014].

Czajka, J.L. (2009). SIPP data quality. Appendix A in *Reengineering the Survey of Income and Program Participation*. National Research Council. Washington, DC: The National Academies Press.

Czajka, J.L. and Denmead, G. (2012). Income measurement for the 21st century: Updating the current population survey. Washington, DC: *Mathematica Policy Research*. Available: http://www.mathematica-mpr.com/~/media/publications/PDFs/family_support/income_measurement_21 century.pdf [July 2014].

Czajka, J.L., Jacobson, J.E. and Cody, S. (2004). Survey estimates of wealth: A comparative analysis and review of the Survey of Income and Program Participation. *Social Security Bulletin*, 65(1). Available: http://www.ssa.gov/policy/docs/ssb/v65n1/v65n1p63.html [July 2014].

Daas, P.J.H., Ossen, S.J.L., Tennekes, M. and Nordholt, E.S. (2012). Evaluation of the quality of administrative data used in the Dutch virtual census. *Paper presented at the Joint Statistical Meetings*, San Diego, CA (August). Methodology Sector and Division of Social and Spatial Statistics, Statistics Netherland, The Hague.

De Leeuw, E.D. and De Heer, W. (2002). *Trends in Household Survey Nonresponse: A Longitudinal and International Comparison*. R.M. Groves, D.A. Dillman, J. L. Eltinge and R.J.A. Little, eds. Survey Nonresponse, 41-54. New York: Wiley.

Duncan, J. W. and Shelton, W. C. (1978). *Revolution in United States Government Statistics 1926–1976*. Office of Federal Statistical Policy and Standards, U.S. Department of Commerce. Washington, DC: Government Printing Office.

Eurostat. (2000). Assessment of the quality in statistics. *Doc. Eurostat/A4/Quality/00/General/Standard report*. Luxembourg (April 4-5). Available: http://www.unece.org/fileadmin/DAM/stats/documents/2000/11/metis/crp.3.e.pdf [July 2014].

Fixler, D. and D.S. Johnson (2012). Accounting for the distribution of income in the U.S. National Accounts. *Paper prepared for the NBER Conference on Research in Income and Wealth*, September 30. Available: http://www.bea.gov/about/pdf/Fixler_Johnson.pdf.

Fricker, S. and R. Tourangeau (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, 74(5), 935-955.

Griffin, D. (2011). Cost and workload implications of a voluntary American community survey. *U.S. Census Bureau,* Washington, DC (June 23).

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(9), 861-871. Special 75th Anniversary Issue.

Groves, R.M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2), 167-189.

Harris-Kojetin, B. (2012). *Federal Household Surveys*. Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey, Second Edition, M. J. Anderson, C.F. Citro and J.J. Salvo, eds, 226-234. Washington, DC: CQ Press.

Heckman, J. J. and LaFontaine, P.A. (2010). The American high school graduation rate: trends and levels. *NBER Working Paper 13670*. Cambridge, MA, National Bureau of Economic Research. Available: http://www.nber.org/papers/w13670 [July 2014].

Hendriks, C. (2012). Input data quality in register based statistics-The Norwegian experience. Proceedings of the *International Association of Survey Statisticians-JSM 2012*, 1473-1480. Paper presented at the Joint Statistical Meetings, San Diego, CA (August). Statistics Norway, Kongsvinger, Norway.

Hicks, W. and Kerwin, J. (2011). Cognitive testing of potential changes to the Annual Social and Economic Supplement of the Current Population Survey. *Report to the U.S. Census Bureau,* Westat, Rockville, MD (July 25).

Holt, D.T. (2007).The official statistics Olympics challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician*, 61(1, February), 1-8. With commentary by G. Brackstone and J.L. Norwood.

Horrigan, M.W. (2013). Big data: A BLS perspective. *Amstat News*, 427(January), 25-27.

Hoyakem, C., Bollinger, C. and Ziliak, J. (2014). The role of CPS nonresponse on the level and trend in poverty. *UKCPR Discussion Paper Series*, DP 2014-05. Lexington, KY: University of Kentucky Center for Poverty Research.

Iwig, W., Berning, M., Marck, P. and Prell, M. (2013). Data quality assessment tool for administrative data. Prepared for a subcommittee of the *Federal Committee on Statistical Methodology*, Washington, DC (February).

Johnson, B and Moore, K. [no date]. Consider the source: Differences in estimates of income and wealth from survey and tax data. Available: http://www.irs.gov/pub/irs-soi/johnsmoore.pdf [July 2014].

Keller, S.A., Koonin, S.E. and Shipp, S. (2012). Big data and city living - what can it do for us? *Statistical Significance*, 9(4), 4-7, August.

Kennickell, A. (2011). Look again: Editing and imputation of SCF panel data. *Paper prepared for the Joint Statistical Meetings*, Miami, FL (August).

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group [now Gartner] Research Note*, February 6. See: http://goo.gl/Bo3GS [July 2014].

Manski, C.F. (2014). Communicating uncertainty in official economic statistics. *NBER Working Paper No. 20098*. Cambridge, MA: National Bureau of Economic Research.

McGuire, T., Manyika, J. and Chui, M. (2012). Why big data is the new competitive advantage. *Ivey Business Journal* (July-August).

Meyer, B. D. and Goerge, R.M. (2011). Errors in survey reporting and imputation and their effects on estimates of Food Stamp Program participation. Working Paper. *Chicago Harris School of Public Policy*, University of Chicago.

Meyer, B.D., Mok, W. K.C. and Sullivan, J.X. (2009). The under-reporting of transfers in household surveys: Its nature and consequences. *NBER Working Paper No. 15181*. Cambridge, MA: National Bureau of Economic Research.

Moore, J.C., Marquis, K.H. and Bogen, K. (1996). The SIPP cognitive research evaluation experiment: Basic results and documentation. *SIPP Working Paper No. 212*. U.S. Census Bureau, Washington, DC (January). Available: http://www.census.gov/sipp/workpapr/wp9601.pdf [July 2014].

Morganstein, D. and Marker, D. (2000). A conversation with Joseph Waksberg. *Statistical Science*, 15(3), 299-312.

Mule, T. and Konicki, S. (2012). *2010 Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Housing Units in the United States*. U.S. Census Bureau, Washington, DC.

National Research Council (1995). *Measuring Poverty: A New Approach*. Washington, DC: National Academy Press.

National Research Council (2004). *The 2000 Census: Counting Under Adversity*. Washington, DC: The National Academies Press.

National Research Council (2010a). *Envisioning the 2010 Census*. Washington, DC: The National Academies Press.

National Research Council (2010b). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.

National Research Council (2013a). *Measuring What We Spend: Toward a New Consumer Expenditure Survey*. Washington, DC: The National Academies Press.

National Research Council (2013b). *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: The National Academies Press.

National Research Council (2013c). *Principles and Practices for a Federal Statistical Agency*. Washington, DC: The National Academies Press.

Nelson, N. and West, K. (2014). Interview with Lars Thygesen. *Statistical Journal of the IAOS*, 30, 67-73.

Passero, B. (2009). The impact of income imputation in the Consumer Expenditure Survey. *Monthly Labor Review* (August), 25-42.

Prell, M., Bradsher-Fredrick, H., Comisarow, C., Cornman, S., Cox, C., Denbaly, M., Martinez, R.W., Sabol, W. and Vile, M. (2009). Profiles in success of statistical uses of administrative records. Report of a subcommittee of the *Federal Committee on Statistical Methodology*, U.S. Office of Management and Budget, Washington, DC.

Shapiro, G.M. and Kostanich, D. (1988). High response error and poor coverage are severely hurting the value of household survey data. *Proceedings of the Section on Survey Research Methods*, 443-448, American Statistical Association, Alexandria, VA. Available: http://www.amstat.org/sections/srms/Proceedings/papers/1988_081.pdf [July 2014].

Steeh, C.G. (1981). Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly*, 45, 40-57.

Tourangeau, R. (2004). Survey research and societal change. *Annual Review of Psychology*, 55, 775-801.

U.S. Office of Management and Budget. (2014). *Guidance for Providing and Using Administrative Data for Statistical Purposes*. Memorandum M-14-06. Washington, DC.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Woodward, J., Wilson, E. and Chesnut, J. (2007). *Evaluation Report Covering Facilities - Final Report. 2006 American Community Survey Content Test Report H.3.U.S.* Census Bureau. Washington, DC: U.S. Department of Commerce. January.

Zelenak, M.F. and M.C. David (2013). *Impact of Multiple Contacts by Computer-Assisted Telephone Interview and Computer-Assisted Personal Interview on Final Interview Outcome in the American Community Survey*. U.S. Census Bureau, Washington, DC.