Navigating the Shoals of Self-Reporting:  
Data Collection in U.S. Expenditure Surveys since 1920

Thomas A. Stapleford  
University of Notre Dame  
October 2011

(Published with minor modifications in History of Political Economy, 44 (supplement, 2012): 160 – 181).

This essay examines the evolution of data-collection practices in American expenditure surveys over much of the twentieth century. Economists conducting expenditure surveys faced one fundamental concern: their success hinged upon the cooperation of interviewees (typically housewives) and the reliability of their testimony. Investigators recognized that both dependencies posed serious problems, and they struggled to devise effective solutions. I argue that over the course of the twentieth century, the methodology of U.S. expenditure surveys evolved in a non-teleological way, as new approaches led the BLS to adopt previously rejected techniques and to abandon other strategies that were formerly of central importance. In charting this history, the essay reveals the many challenges and complexity of gathering a seemingly straightforward form of economic data.

Introduction

If there is one domain of economics grounded on direct observations, it would seem to be economic statistics. For more than a century, quantitative data have provided raw material for debates about political economy and economic theory. But describing economic statistics as “observations” is misleading: strictly speaking, many economic data – unemployment rates, wages, national income, consumer expenditures – are not based on observations at all. Rather, they are derived from self-reporting by the original participants, whether businesses or households.

Alongside the familiar methodological quandaries that accompany the practice of observation in general (such as the role of theory in guiding observation or the concern that observational practices may alter the phenomena under study), self-reporting raises a new dilemma: How can we know that self-reported data are accurate, i.e., that the data match what the survey’s designers would have recorded had they been directly privy to the transactions in question?
Deliberate deception is a serious issue, especially when participants have incentives to provide misleading information (e.g., during surveys of workplace safety or employment practices). Many early surveys wrestled mightily with deception through the early twentieth century, especially from employers suspicious of government intrusion. Deception was not the only threat to accurate self-reporting, however: potential subjects might refuse to participate because of privacy concerns or lack of time; they might misunderstand the queries and hence answer incorrectly; or they might simply lack the knowledge needed to provide accurate information.

This essay examines how economists have responded to the challenges posed by self-reporting in economic statistics. My focus is on the major surveys of household income and expenditures administered by the U.S. government during the twentieth century, especially those coming from the U.S. Bureau of Labor Statistics (BLS). Statisticians and labor activists began conducting expenditure surveys in the late nineteenth century in order to document the typical living conditions of working-class families. In the 1930s, however, federal expenditure surveys expanded to encompass a new goal that soon came to dominate the projects: providing empirical data for the study of consumer demand (Stapleford 2007). Over the next few decades, this shift prompted a two-fold change in the focus of BLS expenditure surveys: from the working-class to the coverage of all American consumers and from families to all households (including single-individuals).

Expenditure surveys provide a valuable case study in the problems of self-reporting because, unlike businesses, households do not employ accountants and rarely keep even minimally-detailed

---

1 See the frequent discussions of this issue at the annual conventions of the National Association of Bureaus of Labor Statistics, especially the tenth convention (1894). For concerns about deception in the twentieth century, see Marcel Boumans’s essay on Morgenstern, this volume.

2 Since all major surveys of household expenditures conducted by the U.S. government also covered household income, I will refer to these studies simply as “expenditure surveys,” with the income component being implied.
records. As a result, expenditure surveys face numerous challenges: households lack the expertise that can ameliorate the risk of miscommunication with investigators seeking financial data; they are forced to rely more heavily on memory; and responding to inquiries can take a substantial amount of time (because the requisite information is not readily at hand.) The efforts of BLS statisticians to wrestle with these obstacles forms a particularly attractive topic because the bureau operates the most prominent expenditure surveys in the United States, has a long tradition of running such surveys (see Table 1), and is far more open about its methods than are private firms devoted to market research.

**The evolution of data-collection methods: an overview**

A simple history of BLS survey procedures does not present a clear teleology: some practices that were once rejected as inefficient or misleading were later re-adopted, while other long-favored strategies were subsequently critiqued. To some degree, this lack of teleology reflects the changing context for expenditure surveys: as I discuss in the conclusion, both funding constraints and changes in the structure of households and the consumer economy have necessarily shaped BLS survey methods. Yet it also reflects conceptual shifts in which new approaches to survey research came to supplement or even supplant earlier practices.

The first change was the gradual rationalization of survey procedures, i.e., an increasing reliance on highly standardized rules. From the late nineteenth century into the early decades of the twentieth, the BLS confronted the difficulties of self-reporting by relying on the personality, charisma, and social skills of its field agents to elicit the most accurate answers from respondents. By the early 1930s, however, the bureau had followed other federal agencies in shifting its focus to standardized rules and formal procedures, encompassing both survey design and the training and oversight of field agents. In a classic pattern of Weberian rationalization, BLS staff began to treat
survey methodology as an object for scientific observation and experimentation in its own right, taking these results and feeding them back into bureau practices. In this framework, each survey became a double-experiment, yielding information about both economic activity and the survey’s methodology. The goal of this process was to elicit the most accurate and complete responses for a given household, an objective that in principle could be reached after a sufficient number of refinements.

By the 1960s, however, BLS survey methodology was in crisis as conflicting results and the inherent limitations of self-reporting suggested that no single set of procedures would result in consistently accurate and complete responses for a full range of households. Rationalization, it seemed, had reached its limits. At this stage federal statisticians developed a simple yet (in context) radical solution: rather than collecting complete expenditure data from each household in the sample, the survey would be split into two components, with each household reporting only part of its annual expenditures. While a logical step, this change forced statisticians to abandon certain tests for internal coherence and consistency that had previously been viewed as critical tools for filtering out sloppy or inaccurate survey responses. I argue that the new procedures therefore reflected a growing confidence in the bureau’s research on survey methodology. In essence, the BLS trusted that research would reveal and eliminate systematic response errors while residual (random) inaccuracies would disappear when combined into average values. As a result, survey responses no longer had to be discarded if they were incomplete or contained suspected mistakes; indeed, the BLS even began to “correct” incomplete survey responses based on typical patterns.

**Rationalizing household interviews**

During the first several decades after the founding of the BLS in the late nineteenth century, the agency treated its field agents as highly skilled workers who could be granted a relatively substantial
amount of autonomy. The success of survey research thus rested on the agents’ charisma, social intuition, and previous experience in fieldwork. This labor model encountered problems, however, as the geographic scope of BLS surveys expanded, and especially when the bureau had to hire large numbers of temporary workers for intermittent projects. Accordingly, the BLS began to impose tighter supervision and oversight on its agents in the early decades of the twentieth century (Stapleford 2010). Nonetheless, the agency retained its emphasis on natural abilities and social skills as the key to successful interviews for household surveys.

The persistence of that tradition is evident in the description of an ideal field agent for expenditure surveys given by BLS commissioner Ethelbert Stewart in the mid-1920s. According to Stewart, “a careful high school education and mathematical accuracy or ordinary intelligence as applied to the mathematical side,” would be sufficient for a field agent. The keys to success lay elsewhere: “Kindness and a knack at making friends, ability to state her mission in perfectly simple, impressive and appealingly argumentative language, and pleasing personality are the basic requirements of a good agent, when it comes to getting the basic information from the housewife herself.” By use of her persuasive powers, a good agent could turn the housewife into a strong ally, making her “really interested for her own sake in working out an itemized statement of the quantities and cost of things she consumes in her household.”

Likewise, draft instructions for an uncompleted expenditure survey encouraged field agents to rely on their social skills and intuition to pry information from recalcitrant subjects:

The agent should work on the assumption that there are no facts called for on the schedule which cannot be determined if properly sought after. To this end she should use patience, perseverance, tact, and skillful and ingenious questioning. She should determine by

---

3 Ethelbert Stewart to Secretary of Labor, 25 November 1927, folder 2, box 1, Correspondence of the Commissioner with the Secretary of Labor and Units of the Labor Dept., 1925-1929, Records of the BLS.
experimentation the best order in which to take up questions so as to avoid tiring the housewife so quickly that the schedule cannot be completed accurately.4

As the gendered pronouns in these descriptions implied, the BLS typically hired women for expenditure surveys in the early twentieth century, in part (one suspects) because they were thought to be more effective dealing with fellow women (housewives) and in part because the emphasis on “kindness,” “patience,” and conversational acumen resonated with twentieth-century conceptions of feminine characteristics.5

Even during this period, of course, BLS staff recognized that proper technique could improve survey results, and indeed that some social skills could be explicitly taught to agents. For example, the draft instructions reminded agents that “in questioning a family as to surplus and deficit, it is sometimes advantageous to avoid the use of the words, ‘Surplus, deficit, savings, and debt,’ and to approach the subject in a more tactful way by asking such questions as, ‘Were you able to put aside or invest anything for a rainy day?’, ‘Were you able to meet all of the year’s expenses with the year’s earnings?’, etc.”6 Likewise, the bureau instructed agents to be cognizant of typical prices and quantities consumed for basic household goods in the cities under study and “to call [the housewife’s] attention in a kindly way to what seems to be an unusual statement.” Yet even here, the judgment and social graces of the field agent took a central role: Ethelbert Stewart insisted that agents who encountered apparent discrepancies should utilize “common sense, but above all kindness,” since “it very often happens that these unusual statements are true.”7

---

4 “Instructions, Family Budgets, July 1931,” pp. 2-3, Instructions For Future Study--Developed from 1918 Survey, Box 1, Records of the Consumer Expenditure Survey Program and Predecessors, 1933-1971, Records of the BLS.
5 That trend continued in later years as well, though the bureau relied on men for nighttime interviews and visits to “less desirable areas of the cities” (Lamale 1959, 64).
6 “Instructions, Family Budgets, July 1931,” p. 11.
7 Stewart to Secretary of Labor, 25 November 1927.
By contrast, during the major expenditure surveys conducted by the BLS in the mid-1930s, the previously explicit emphasis on “kindness,” “pleasing personality,” and social acumen largely disappeared. Though charisma and conversational skill were probably factors in hiring decisions and undoubtedly affected the success of various agents in the field, they were no longer discussed in BLS documents. Job descriptions during the 1930s emphasized professional qualifications (“preferably…should have college training in economics or sociology and research experience or comparable business experience….“); the only social skill listed was a vague reference to “good judgment in making contacts.” Likewise, for the next major survey (1950) staff screened applicants through a standardized federal aptitude test that measured “(1) general intelligence; (2) verbal ability; (3) numerical ability; and (4) clerical perception.” Field supervisors did interview prospective agents in part to assess their “personality traits,” but the latter referred primarily to factors that might distort the results, such as “class consciousness” or a tendency “to become emotionally involved in the personal lives of the respondents” (Lamale 1959, 63).

Instead of the charisma and “ingenious questioning” of expert field agents, BLS staff members from the 1930s onward placed their faith in proper survey design, detailed procedural guidelines, and close oversight of field agents. As far as possible the entire survey process would be tightly specified and regulated by the staff in Washington, who would plan and administer the investigation based on the best, empirically-tested strategies for maximizing accurate returns. The process paralleled the kind of Weberian rationalization that occurred in manufacturing, where Taylorist disciples codified procedures as a way of standardizing production, often leading to a de-skilling of the workforce in the process. Not surprisingly, the drive towards rationalization was partly

---

9 For a more detailed examination of this transformation in the BLS during the 1930s, see Stapleford (2003, 322-358).
a response to the labor problems created by intermittent, large-scale surveys. Such projects required high numbers of temporary workers and (just as occurred with the U.S. census) had been leading the BLS to increase its oversight of field agents since the turn of the century (Stapleford 2010). The New Deal accelerated this process across the federal bureaucracy as administrators used large, short-term statistical surveys as white-collar relief projects in the early years of the Roosevelt administration, creating more than 2.5 million temporary positions for field agents in just a few years and producing what Emmanuel Didier has called the “industrialization” of statistical field agents (Didier 2011).

Yet the renewed push to rationalize BLS field work in the 1930s also reflected a change in the ethos of the bureau’s top staff. Prior to the New Deal, the BLS had been run largely by self-styled “practical statisticians” who had little advanced academic training and who often regarded effective data collection as an art learned through practical experience. The arrival of the Roosevelt administration brought a wholesale overhaul of the bureau’s top staff, with the older generation replaced by a young cadre of social scientists (primarily economists) who shared the common belief that empirically-grounded procedures offered the key to a more rigorous, scientific knowledge about American society (Stapleford 2009, 145-169). Rationalization offered the alluring promise of ever-increasing accuracy and detail made possible by refinements in method, refinements that were largely under the oversight of the top staff who designed the survey forms, devised the overall methods for each investigation, and wrote the instruction manuals. Rules and procedures became a focus in part, in other words, because they were something that managerial staff could control.

Part of the new strategy involved crafting formal guidelines and training procedures for interview techniques that (ideally) would compensate for any deficiencies in a field agent’s personality or social expertise. Agents participating in a housing survey in the late 1940s, for example, received a three-page sheet describing “Interview Technique[s].” Some were basic social
skills, adapted for an effective sales pitch: “Look at respondent”; “Talk slowly”; “Don’t pause between sentences”; “Sell, don’t order.” Agents were told to take control of the conversation and to have “a ready answer for every objection.” To that end, the staff provided seven common objections with responses, for example: “‘Not interested’--This survey is important,” or “‘Too busy’--It only takes a few minutes.” Likewise, the training regimen for field agents in the 1950 BLS expenditure survey included a battery of practice interviews plus a cartoon film on “Some Helpful Guides for Interviewing” (Lamale 1959, 63, 65).

Yet the effort to rationalize charisma and conversational acumen was never entirely successful, in the BLS or elsewhere. After several decades of research by psychologists and other social scientists, two experts in the 1970s concluded that although “rapport” was universally-regarded as critical to the success of any survey interview, the concept could not be rigorously defined for experimental purposes, and hence could not be studied effectively (Goudy and Potter 1975). Likewise, another scholar acknowledged that “though rapport-building techniques are an essential part of the data-collection process,” they were also inherently limited. Accordingly, “many investigators prefer to seek more mechanical methods of controlling extraneous variables that do not depend so heavily on ad hoc procedures and unspecifiable skills” (Scott 1968, 239).

For the BLS, as for other survey researchers, these limitations led to a greater emphasis on the aspects of the survey process that could be controlled, tested, and refined: the frequency of interviews, procedures for handling non-responses, the structure of the survey form and the interview itself, the phrasing of questions, strategies for evaluating the accuracy of responses, and so forth. As a result, the BLS began to spell out proper survey procedures in immense detail. Whereas draft instructions to field agents from the 1920s staff totaled only 28 pages, the 1930s version was

---

nearly three times as long, and that was only one of an extensive range of new manuals produced by 1935. Meanwhile, the staff created an extensive system for training and monitoring field agents, including exams, closely-guided reviews of completed “schedules” (survey forms), and regular “check-interviews” in which supervisors revisited homes and re-interviewed housewives in order to double-check an agent’s report.\textsuperscript{11} For the 1950 expenditure survey, basic training for field agents took a full seven days and included an assortment of “audio-visual aids” to supplement the extensive manuals (Lamale 1959, 65). Not surprisingly, a 1956 study in Britain (cited approvingly by BLS staff) concluded that field agents who returned schedules with the smallest response errors were those who had the ability and willingness to follow complicated rules; by contrast, poor field agents manifested “an impatience with a formal system of collecting facts” (Cole 1956, 58; cf. Lamale 1959, 28). Indeed, following the rules had become the major desideratum: as the BLS explained, “the techniques for hiring and training the interviewers” for the 1950 expenditure survey “were designed to reduce the interviewer variability as much as possible” by ensuring that agents would follow instructions precisely (Lamale 1959, 29). In 2003, the rationalization of survey work reached its next natural terminus with the introduction of the Computer Aided Personal Interview, a software package that permits direct data entry and guides field agents through the interview process (U.S. Bureau of Labor Statistics 2007, 3).

Rationalization—with its emphasis on rules and procedures—led BLS staff to envision expenditure surveys as double-experiments: studies of the economy and studies of how to collect data. If the accuracy of a survey depended solely on the formal procedures and tools used to gather data (survey forms, instructions to field agents, etc.), then one could test the efficacy of different procedures experimentally. Though BLS staff had always learned from past experiences and used

previous surveys as models, that viewpoint became far more formal and deliberate after 1933. Thus, for example, the 1940s staff described innovative survey schedules as “experimental in character” and explained that “in a very real sense, each survey constitutes a methodological test for the one that follows it” (Brady and Williams 1944, 341, 333). In turn, this view of surveys as methodological experiments led to new practices. Previously, BLS staff rarely archived any survey documents beyond a few blank schedules and (occasionally) the completed forms for major surveys (such as the 1917-1918 national expenditure survey). The 1930s brought a radical leap in archival record-keeping, as staff preserved instruction manuals, memos to and from the field, minutes and memos from the planning and tabulation processes, a broad range of sample forms, and extensive documentation of methodological discussions. The organization and labeling of these materials strongly suggests that they were preserved as references for future expenditure surveys. Not surprisingly, BLS publications on the next major expenditure survey (1950) included a volume on methodology in which BLS staff member Helen Humes Lamale provided a historical overview of expenditure survey methodology, including detailed information on previous BLS projects (Lamale 1959, 1-40, 174-233). Indeed, the very existence of Lamale’s monograph reflected the rise of rationalization: alongside other volumes in the series that sketched a history of family incomes and expenditures, the BLS offered a history of the surveys themselves, one that highlighted the refinement of survey procedures and the current state-of-the art.

The most striking innovation that accompanied rationalization, however, was the rise of small-scale, experimental surveys that were designed primarily to test methodology rather than collect data per se. In the late 1920s, the Bureau of Home Economics (BHE) in the Department of Agriculture ran a small test survey intended to compare the accuracy of family “accounts” or diaries

---

12 Records from the 1930s, for example, were housed in a separate sub-collection, eventually titled “Records of the Consumer Expenditure Survey Program and Predecessors, 1931-1971.”
(when families recorded purchases on a standardized form on a daily basis) and responses from an interview (when a field agent asked families to recall what they had purchased over a set period, ranging from one month to a year) (Williams 1932, 3). When the BHE’s supervisor of family expenditure surveys, the economist Faith M. Williams, moved to the BLS in 1934, she continued to explore the possibilities of experimental surveys. As a result, the BLS ran several independent experiments on data collection techniques during the 1940s and began incorporate carefully-designed “pilot surveys” into every major project so that staff could test “the phrasing of the questions, the selection of details to be recorded, and the inclusion of probes and check questions” (Brady and Williams 1944, 332-333; Lamale 1959, 6, 19). In later decades, the BLS established a Statistical Methods Division to oversee research on data collection, and by the early 1990s this had been supplemented by a Collections Procedures Research Laboratory that performed “cognitive testing” of questions and survey forms on small groups of subjects (Dippo 1987; Jacobs and Shipp 1993, 71).

The results and crisis of mid-century methodological research

Results

BLS research on survey methodology during the middle third of the twentieth century seemed to converge on several major results. The first and perhaps most durable finding was that human memory is more effective when given specific prompts. BLS experiments during the early 1940s revealed that checklists (where subjects were asked about specific items) led to more self-reported purchases than “free-listing” (when subjects were asked to write down purchases in general categories); likewise, more detailed lists tended to produced correspondingly more complete reports of expenditures (Brady and Williams 1944, 331-332; Lamale 1959, 18-19). Accordingly, BLS survey forms rapidly ballooned. The standard “schedule” for the 1917-1919 BLS expenditure survey was 11
pages, which expanded to 20 pages in the 1934-1936 version. By 1950, the length had jumped to 59 pages, and by the 1990s it was over 100 pages (Lamale 1959, 189, 208, 261-304, 321-335; Jacobs and Shipp 1993, 62).

The second major conclusion, albeit one that proved far more controversial, was that expenditure surveys should be based primarily on an interview format (in which field agents asked subjects to recall purchases over different periods, usually the previous week and the previous year) rather than the “diary” method (in which subjects recorded purchases as they occurred). U.S. statistical agencies had a long tradition of favoring interviews for expenditure surveys dating back to the first federal surveys in the 1890s (e.g., U.S. Commissioner of Labor 1891). However, the diary method was much more popular outside of the United States and was also a common component of American market research (International Labour Office 1939, 671; Lamale 1959, 4, 6). Not surprisingly, in light of this divergence, survey format became a focus for federal research in the 1930s and 1940s (including the BHE’s first survey experiment, discussed above).

By 1934, BLS staff had developed a provisional view that was reinforced by subsequent research over the next two decades. Diaries, Faith Williams conceded in 1935, “would seem to be the logical one to use, as a record of events made day by day as they occur should be more accurate than a report of these same events from memory several months later” (Williams and Zimmerman 1935, 45) Nonetheless, that abstract advantage was undercut by critical weaknesses in practice. First, statisticians wondered if the need to record expenditures would alter family purchasing habits, a speculation that was apparently confirmed by research showing that families asked to track food expenditures tended to report total costs that were lower than those obtained through interviews (Cornfield 1942, 67-68). Second, American statisticians worried that the extra burden of keeping diaries would skew the sample of families participating in a survey. To provide a complete picture of expenditures, surveys needed to cover a lengthy period; the International Labour Office, for
example, recommended gathering “daily records of income and expenditure…for a period of 12 months.” But how many families would be willing to undertake such a project, and would they be representative of Americans as a whole? Williams was skeptical (Williams and Zimmerman 1935, 46), and again, subsequent research suggested that diary-based surveys suffered from very high refusal rates and samples that overrepresented families with (relatively) higher incomes, more education, and greater community involvement (Lamale 1959, 7-12). Finally, diaries seemed to promise higher costs. Not only would tabulation be more expensive (because of the need to aggregate numerous daily entries), but an early BHE study demonstrated that accurate results required “frequent visits to the family or a continuous editing and correspondence,” thereby leading Williams to conclude that the costs would be substantially higher than for a comparably-sized survey based on interviews (Brady and Williams 1944, 329-330).

Despite these concerns about diaries, Williams and her colleagues also recognized the potential weaknesses of an interview that required participants to recall their annual expenses in a full range of categories (including food, clothing, shelter, furnishings, appliances, medical care, insurance, and so forth). Accordingly, the BLS adopted a hybrid procedure for its major surveys from the mid-1930s through the early 1960s. The main survey remained an interview that covered annual expenses across the full range of expenditure categories. In addition, however, the bureau asked respondents to record expenditures during the last seven days for a subset of frequently-purchased goods (notably food, drugs, and personal care items), and the agency occasionally supplemented that portion with one or more weekly expenditure diaries recorded by the family (cf. Lamale 1959; U.S. Bureau of Labor Statistics 1971).

Finally, continued experience with expenditure surveys in the years after 1930 (with a concerted focus on technique) revealed a number of systematic biases in subjects’ responses. Many of these biases became apparent as the bureau sought to evaluate the accuracy of their data by cross-
checking individual responses for internal consistency and comparing aggregate results with data from other sources. For example, during the 1917-1919 survey, the BLS told field agents that if the total income and total expenditures (including savings) did not match for a given respondent, then “the figures of one item or the other must be revised” before the schedule was submitted. As Williams put it dryly, “this procedure seems to encourage ‘adjustments’ in the figures provided by the family to a dangerous point” (Brady and Williams 1944, 340). For the bureau’s 1934-1936 survey, agents were asked to re-interview families with substantial imbalances between income and expenditures, but not to force the schedule the matter; instead, the BLS just discarded schedules where the final imbalance between overall money receipts and money expenditures exceeded 5.5%.

Likewise, housewives had to give both an annual expenditure estimate for food and a detailed account of the last week; 52 x the one-week total had to come within 15% of the annual estimate for the schedule to be tabulated.

These new procedures allowed BLS statisticians to begin studying patterns in schedule imbalances, which proved to have certain tendencies. Overall, Lamale could report in 1959, “average family income is usually understated; average family expenditures are also understated but somewhat less; and average savings are greater or dissavings less than shown by the reported changes in assets and liabilities” (Lamale 1959, 22). Comparing BLS data with aggregate data from other sources (such as tax returns or sales figures from the Department of Commerce) gave the agency confidence that these imbalances represented true errors rather than legitimate trends in income / expenditure patterns. Such comparisons only became possible after the 1930s, when the BLS expanded its expenditure survey program to encompass all urban consumers and partnered with the Department

---

13 “Instructions to Field Agents,” pp. 5-6, 76.
14 “Instructions to Editors,” February 1936, p. 8, Instructions--1934-36, Box 1, Records of the Consumer Expenditure Survey Program and Predecessors, 1933-1971, Records of the BLS.
of Agriculture to collect data on rural families as well, thereby providing comprehensive coverage that could readily be extrapolated to national aggregates.

These comparisons allowed the identification of more specific problems. Total expenditures ran (on average) about 10 percent less than Department of Commerce figures, with the largest shortfalls coming in food and clothing (especially for luxury goods such as jewelry, where survey results were “strikingly low”). These underestimations were partially offset by higher survey results for housing, fuel, light, and refrigeration (though the BLS had no solid explanation for the discrepancy). Most of the income shortfall, meanwhile, arose from undercounting interest and dividends. Finally, the effort to survey all Americans revealed systematic sampling problems: families at both ends of the income spectrum (but especially the higher end) were much more likely to refuse to participate in the survey (Brady and Williams 1944, 326-328, 341-343).

In principle, most of these problems could be addressed by further refinements in procedures. Oversampling could offset higher refusal rates among certain population segments, for example, and staff hoped that more detailed survey questions would reduce underreporting by improving recall (e.g. Brady and Williams 1944, 328). For suspected cases of “intentional bias,” as when consumers understated expenditures on alcohol or luxury items such as jewelry, the bureau developed new interviewing strategies, such as asking when an item from that category was last purchased and what price was paid. Total expenditures were then estimated by extrapolating from this response, leading to higher totals than obtained through open-ended questioning (Lamale 1959, 20). Given these possibilities, there seemed no reason to doubt that researchers would continue to improve the accuracy of individual, household reports well into the future.

A growing crisis
By the late 1960s, however, any such optimism had dissipated. In part, some of the mandates developed through rationalization pulled the bureau in opposite directions. For example, as we have seen, the most basic lesson of methodological experiments was that the accuracy of individual returns increased when interviewers asked more detailed questions. However, more detailed questions also required more time, potentially reducing participation rates or skewing samples. Already in the 1930s, one experienced market researcher warned that the bureau’s survey questions were “so detailed, so numerous, and so inquisitorial that no ordinary person will care to answer them;” indeed, he suggested, only “abnormal” people would even know the relevant information (White 1936, 23). By 1950, when the average interview took eight hours to complete, such fears seemed increasingly plausible. Responding to “widespread concern about the length of the interview” and its potential consequences for response rates and “inaccuracy…because of fatigue,” the BLS sought to shorten the interview time for the 1960 survey by eliminating duplication and consigning some peripheral questions to other surveys. Nonetheless, the drive to improve accuracy led to expansions in other areas. As a result, the net reduction was minimal, with the average interview still lasting just under seven hours (U.S. Bureau of Labor Statistics 1971, 6, 17, 24).  

A similar conflict occurred over the bureau’s various requirements for internal consistency within a given schedule, such as the balance between total income and expenditures or the comparison of annual and weekly food estimates. The 1930s staff had used these tests to gauge the accuracy of results, using them to trigger second interviews and excluding schedules with imbalances above certain levels. However, by emphasizing the importance of consistency, the bureau also risked pressuring agents and families to make artificial adjustments merely to conform to expected values.

---

15 For the 1960 survey, the bureau insisted that tests indicated no substantial increase in refusal rates between an interview requiring four hours and one requiring seven hours; still, interview time clearly remained a concern. Today the typical interview for the BLS consumer expenditure survey lasts just over one hour (U.S. Bureau of Labor Statistics 2007, 2)
standards. Already in 1936, a young Milton Friedman (working for a federal agency overseeing a major expenditure survey) had warned supervisors that his visits to field offices had revealed “a danger of a tendency to ‘force a balance’” in order to make schedules acceptable. 16 Indeed, after a pilot study in 1949, the bureau itself highlighted “the danger of placing too much emphasis on a balancing criterion,” and it weakened the criterion for the 1950 study, using imbalances as prompts for revisiting families but not as an eligibility requirement (Lamale 1959, 25). That policy mitigated the threat of over-enthusiastic editing, but at the cost of what earlier staff members had considered to be a basic safeguard. For the 1960 survey, the BLS retained the same general procedure but added a “professional review” of all schedules that in which highly-trained staff members scrutinized schedules for internal consistency and attention to detail. Most schedules were accepted; some were rejected; and a few (roughly one percent) were “improved” when reviewers felt that they could clearly identify missing items (U.S. Bureau of Labor Statistics 1971, 29-33, esp. 33). Balancing criteria thus hovered uneasily between being essential guides to accuracy and sources of distortion.

The most serious issues, however, involved the growing evidence that a single interview of a household (the bureau’s preferred method) 17 might not be able to capture a complete record of expenditures. The BLS, for example, acknowledged that different goods and services were best captured by recall periods of different lengths: annual, monthly, or weekly (Lamale 1959, 19). Yet extrapolating annual totals from monthly or weekly estimates required either assuming that there was no seasonal variation or undertaking continuous sampling over an entire year. Meanwhile, multiple studies during the 1950s and 1960s illuminated new problems such as telescoping – a phenomenon (common with large expenditures) in which consumers mistakenly attributed

---

16 Milton Friedman to Day Monroe, 22 August 1936, folder 705.1, box 1734, Central Office Records, Records of the National Resources Planning Board.
17 “Single interview” is understood here to also encompass a small set of interviews that all deal with the same time period. Completing a schedule sometimes did require multiple visits to a family over successive days.
purchases from earlier periods to a later period covered by the survey (Neter and Waksberg 1964 esp. 43-49).

Even as statisticians confronted the limitations of interviews, the debate over the value of diaries re-ignited when a major, unpublished study of expenditure survey methodology from the U.S. Bureau of the Census drew attention once again to the contrast between American reliance on annual recall and European use of diaries (Pearl 1968). New experiments on diaries versus recall yielded mixed results, but nonetheless indicated that diaries could have real advantages over recall methods in certain circumstances (Neter 1970, 23-24; Sudman and Ferber 1971). The controversy was made more acute by the context: The study from the Census Bureau came as U.S. statisticians were planning for a new nationwide survey (eventually conducted in 1972-1973) and as the BLS (which had long favored recall interviews) was handing responsibility for administering the survey over to the Census as a cost-saving measure. The result was a sharp dispute that exemplified the broader tensions emerging from BLS research.

Abandoning the complete household survey

Following their existing strategies, U.S. government statisticians sought resolve the impasse by appealing to experiment: in this case, two test surveys that would compare diary and recall methods. Before the results were analyzed, however, the U.S. Office of Statistical Standards (OSS, housed within the Office of Management and Budget) imposed its own solution. The OSS was responsible for coordinating all federal statistics and therefore had final responsibility for approving survey methods, which in turn gave it the power (when necessary) to settle conflicts. In this case, according to two observers, it acted on “a priori statistical and logistical grounds” (Jacobs and Shipp 1993, 67) to develop a novel and elegant strategy that broke with previous tradition.
Rather than continuing to tinker with the standard, annual recall interview, the OSS replaced it with two separate surveys: a quarterly interview panel survey and a diary survey. Households participating in the panel survey were interviewed every three months over a 15-month period and asked about their expenditures. Households in the diary survey recorded their purchases for two weeks. Not only were the two surveys entirely separate, with independent questionnaires and samples, but they each covered only part of the full range of expenditures, with the panel survey emphasizing more expensive, less frequent expenditures while the diary survey focused on smaller, frequently purchased items (U.S. Bureau of Labor Statistics 1978, 2).

The new design alleviated many of the tensions that had come to plague BLS methodology. For example, splitting the survey into two parts made it easier to ask more detailed questions without overwhelming respondents. Moreover, with the quarterly panel survey divided into five visits, the bureau could distribute topics across different visits, allowing some information (such as demographic characteristics, housing costs, appliance purchases, or insurance premiums) to be gathered only on an annual or semi-annual basis. As a result, average time for a single interview plunged to one-and-a-half hours (Jacobs and Shipp 1993, 67) and official response rates rose from roughly 80% in 1950 and 1961-1962 to 89% (U.S. Bureau of Labor Statistics 1978, 4). In fact, the true rise in response rates was substantially higher. The earlier surveys had substituted alternates when the initially-contacted household refused to participate (a procedure that “violated strict probability selection” according to later staff members; Jacobs and Shipp 1993, 65), and the BLS did not count these refusals in the official response rates.18

The flexibility of the panel survey / diary combination also mitigated the weaknesses of the recall method that had been identified over the previous three decades. For example, it allowed the

18 If one did count initial refusals, the response rate for urban households would have been about 75% in 1961-62 per data supplied by the bureau (U.S. Bureau of Labor Statistics 1971, 25)
bureau to design appropriate recall periods (daily, quarterly, semi-annual, etc.) for different goods. Moreover, by using the first interview of the panel survey to take an inventory of the household’s current durable goods, the bureau was able to recognize (and eliminate) much of the “telescoping” that might occur in future interviews (U.S. Bureau of Labor Statistics 1978, 2). Most obviously, of course, the new procedures ended the longstanding conflict over the interviews versus diaries.

The shift to the new format, however, required a fundamental re-conceptualization of the framework for the surveys. The previous approach assumed that the proximal objective was to gather complete expenditure information from individual households which had been randomly selected so as to represent the distribution of American households overall. By contrast, the new strategy dispensed with the focus on individual households; instead, it aimed directly at gathering average expenditures for groups with different socioeconomic characteristics (defined by income, race, geographic location, etc.), a goal that could be reached without having a complete record for any given household.\textsuperscript{19}

In principle, the new approach was perfectly compatible with the overall goals of expenditure surveys. From the nineteenth century onward, statisticians had conducted surveys in order to learn about the characteristic behaviors of different socioeconomic groups, and the emphasis on correlating expenditures with economic and demographic variables had only grown over time (cf. Stapleford 2007). The spending patterns of individual households were of no interest per se; they were only means to the larger end. In that light, one might have expected American statisticians to drop the focus on obtaining complete household records much earlier. In one sense, they did: as early as 1945, BLS staff discussed the possibility dividing a full expenditure survey into

\textsuperscript{19} E.g., one sample of households could provide information on food purchases while a different sample provided information on clothing purchases. As long as the sampling procedures matched, each average could be regarded as truly representative of the population from which the samples were drawn, and therefore the averages could be combined to yield information about average total expenditures on food and clothing.
multiple, independent components, and both the BLS and the U.S. Department of Agriculture conducted small-scale tests of “split-schedule” surveys that same year (Brady and Williams 1944, 337-338; Lamale 1959, 16-18). In both cases, however, the format was rejected, and that decision tells us much about the different conceptual framework that guided American statisticians’ methodological choices at mid-century.

Higher costs were a very important, and very legitimate, concern. (Adding the diaries to the interviews in 1972-1973 almost tripled the number of families visited, and the panel survey required multiple visits over 15 months.) Yet BLS staff also worried about reduced accuracy, and specifically about “the loss of the general criterion of acceptability of a report in the reasonable balance of income, expenditures, and savings” (Brady and Williams 1944, 338). The loss of the balance criterion continued to be mentioned in subsequent years as a significant black mark against the “split-schedule” method, and the results of the test surveys failed to demonstrate significant advantages to the format even as they reinforced fears that the absence of checks for internal consistency would leave results liable to what the Department of Agriculture called “considerable field error” (Lamale 1959, 16-17; U.S. Bureau of Labor Statistics 1971, 3).

Indeed, the adoption of the panel survey / diary combination led directly to the abandonment of balancing criteria altogether. Since each quarterly interview covered only a portion of a household’s expenditures, there was no way to analyze internal consistency by comparing expenditures and income, and subsequent BLS staff members dismissed such analyses as “judgmental” anyway (Jacobs and Shipp 1993, 68). Of course, BLS staff members continued to worry about the accuracy of their results, but they now relied primarily on comparisons with other sources of aggregate data and on experimental research into survey methodology rather than on the

---

20 The full increase was even higher, since the 1972-1973 survey expanded the overall sample size, but this does not seem to be related to the new format.
consistency of individual returns, trusting that the former strategies would provide general guarantees while any real problems in individual returns would have minimal effects on the final tables. As the bureau rather dryly explained, “the data base [for 1972-1973] may contain unusual entries”; nonetheless, the staff remained confident in the verity of its group averages (U.S. Bureau of Labor Statistics 1978, 18).

In essence, therefore, federal statisticians abandoned the balancing criteria (making it possible to split the survey into multiple components) because they now worried less about the accuracy of individual returns. Implicitly, they expected their methodological research to uncover any systematic biases in interview responses while unsystematic (random) errors would cancel each other out in the process of computing an average. The balancing criteria were thus unnecessary. We can see a similar transformation in the bureau’s increased willingness to “correct” incomplete responses on individual schedules. This trend had begun with the 1960-1961 survey, in which staff allocated some expenditures from broad categories into smaller subcategories according to typical patterns if the original family had been unable or unwilling to do so (U.S. Bureau of Labor Statistics 1971, 31). During and after the 1972-1973 survey, this process was dramatically expanded and automated through computer processing (cf. U.S. Bureau of Labor Statistics 1978, 18-19). Today, these post hoc adjustments take three forms: imputing data for “missing and invalid entries” in all fields other than assets (including income and demographic characteristics); allocating expenditures from general categories to more specific subcategories; and classifying expenditures by month in order to calculate calendar-year totals (U.S. Bureau of Labor Statistics 2007, 4). These procedures reflect an increased trust in the verity of BLS averages such that (somewhat paradoxically) one need fret less about the accuracy of the individual returns used to construct those averages. Whereas many of the issues listed above would have once been viewed as signs of sloppy or inaccurate
reporting that might merit disqualification from the survey, these schedules are now readily accepted and “corrected” based on typical patterns.

**Evolution without teleology**

I suggested at the outset of this essay that the history of methodology in BLS expenditure surveys did not present a clear teleology. In part, that judgment rests on the non-cumulative nature of shifts in BLS practices. The balancing criteria that were initially introduced as crucial bulwarks against error-prone schedules, for example, were later discarded as unnecessary and perhaps even as distorting. Likewise, the diary methods that were eschewed in the 1930s and 1940s subsequently became a basic component of the survey program, indeed perhaps the more reliable of the two methods (Jacobs and Shipp 1993, 70). It would be naïve to think evaluations of survey methodology will not be subject to similar fluctuations in the future; conceivably, a new generation of statisticians will decide that the bureau has become overly lenient in its willingness to accept schedules that evince signs of careless reporting.

More generally, though, the non-teleological nature of a history of survey methodology reflects a factor that limited space has prevented me from exploring here: the changing context for survey work. It should be obvious, for example, that funding levels can determine the range of viable methodological options. Though the move to the split panel survey / diary format required a conceptual shift, and only came after a vigorous theoretical debate, it would have been impossible to implement (and thus impossible to consider seriously) had the survey program not received substantially more funding. Likewise, the bureau was fortunate (in a peculiar way) that the 1970s oil shock and subsequent economic upheaval came right after it had completed its 1972-1973 survey, since the poor timing provided the political impetus (and thus the funding) for a continuous survey beginning in 1980 (Jacobs and Shipp 1993, 68-69).
Yet equally important, though more readily overlooked, are changes in the subjects of expenditures, that is, American households and their economic activity. If one considers the radical changes over the last century in household composition, employment patterns, retail practices, financial tools, government interventions (including tax laws), and modes of economic exchange (e.g., credit cards), it becomes evident that even if statisticians have maintained the same abstract goals, they have not been facing the same practical challenges. Why should a set of strategies that worked well in the past continue to function with the same effectiveness in the future? To take but one example, response rates for expenditure surveys (indeed most government surveys) have suffered from continual decline since the 1970s: the response rate for the 1972-1973 survey was nearly 90%; today (with less time-consuming schedules), it is roughly 70 – 75% (U.S. Bureau of Labor Statistics 2007, 5). A history of survey methodology must be non-teleological, in other words, because the surveys themselves are aiming at moving targets. What I have sketched above is therefore only an outline; we await a deeper, richer narrative that will unite the intellectual history of survey methodology with an appropriately dynamic view of American social, economic, and political life.
Table 1: Major Expenditure Surveys of the U.S. Bureau of Labor Statistics

<table>
<thead>
<tr>
<th>Date</th>
<th>Cooperating Agencies</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1888-1890</td>
<td>8,544 families of urban wage-earners in select industries in the U.S. and several European countries</td>
<td></td>
</tr>
<tr>
<td>1901-1902</td>
<td>25,440 families of urban wage-earners and low-salaried clerical workers</td>
<td></td>
</tr>
<tr>
<td>1917-1919</td>
<td>12,096 families of white, urban wage-earners and low-salaried clerical workers</td>
<td></td>
</tr>
<tr>
<td>1934-1936</td>
<td>12,903 families of urban wage-earners and low-salaried clerical workers</td>
<td></td>
</tr>
<tr>
<td>1935-1936</td>
<td>U.S. Bureau of Home Economics</td>
<td>~60,000 families, urban and rural, all occupations and income levels</td>
</tr>
<tr>
<td>1950</td>
<td>12,489 households (single individuals or families) in urban areas, all occupations and income levels</td>
<td></td>
</tr>
<tr>
<td>1960-1961</td>
<td>U.S. Agricultural Research Service</td>
<td>13,728 households in urban and rural areas, all occupations and income levels</td>
</tr>
<tr>
<td>1972-1973</td>
<td>U.S. Bureau of the Census</td>
<td>Partial information (in different forms) from ~70,000 households in urban and rural areas, all occupations and income levels</td>
</tr>
<tr>
<td>1980 ff.</td>
<td>U.S. Bureau of the Census</td>
<td>Continuing survey of households in urban and rural areas, all occupations and income levels. As of 2005, ~45,000 households participate each year.</td>
</tr>
</tbody>
</table>

Archival Sources

Records of the American Federation of Labor, Wisconsin State Historical Society, Madison, Wis.

Records of the National Resources Planning Board, Record Group 187, U.S. National Archives II, College Park, Md.


References


