Volume Title: Improving the Measurement of Consumer Expenditures

Volume Author/Editor: Christopher D. Carroll, Thomas F. Crossley, and John Sabelhaus, editors

Series: Studies in Income and Wealth, volume 74

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-12665-X, 978-0-226-12665-4

Volume URL: http://www.nber.org/books/carr11-1

Conference Date: December 2-3, 2011

Publication Date: May 2015

Chapter Title: Introduction to "Improving the Measurement of Consumer Expenditures"

Chapter Author(s): Christopher D. Carroll, Thomas F. Crossley, John Sabelhaus

Chapter URL: http://www.nber.org/chapters/c12658

Chapter pages in book: (p. 1 – 20)

Introduction

Christopher D. Carroll, Thomas F. Crossley, and John Sabelhaus

As we write in the fall of 2012, many countries (including the United States) are embarking on ambitious multiyear projects to redesign their surveys of household expenditures. In most countries the decision to rethink has been prompted by a sense that existing methods are failing to achieve the surveys' principal objectives, at a time when the importance of those objectives is clearer than ever.

These concerns fit neatly into a broader agenda of improving the measurement of heterogeneity that has been gathering force for a number of years, reflected, for example, in the widely cited work of the Stiglitz-Sen-Fitoussi commission,[1] in the formation of an Organisation for Economic Co-operation and Development (OECD) International Expert Group for the compilation of micro statistics,[2] and in the recent decision by the US Bureau of Economic Analysis to explore constructing "satellite accounts" to account for microeconomic heterogeneity.

Economic theory suggests that a household's spending patterns reflect its economic circumstances better than any other indicator of resources, with the obvious corollary that accurate measurement of households' differences in spending choices would be among the most useful possible tools for understanding economic heterogeneity. This is why the growing concerns

Christopher D. Carroll is chief economist and director of research at the Consumer Financial Protection Bureau. Thomas F. Crossley is professor of economics at the University of Essex. John Sabelhaus is chief of the Microeconomic Surveys Section of the Division of Research and Statistics of the Board of Governors of the Federal Reserve System.

For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see http://www.nber.org/chapters/c12658.ack.

1. Stiglitz, Sen, and Fitoussi (2009).
2. McCall (2012).

about the accuracy of expenditure data are so pertinent to the agenda of measurement of heterogeneity.

This volume brings together work by some of the world's leading experts on measurement of household spending in order to illuminate the difficulties and opportunities that lie ahead for the scholars and statisticians who will be taking up the challenge of producing better data. In broadest terms, the aim of the volume is to provide a knowledge base for agencies and researchers as they design new systems for improving expenditure measurement using household-level data. The volume's sixteen chapters were prepared by economists working on these issues in both academic and government settings, within the United States and in several other countries. (All chapters are based on papers presented at a Conference on Research in Income and Wealth [CRIW] held in Washington, DC, on December 2 and 3, 2011.)

The volume has four main sections. The first provides a framework for analyzing the issues involved in expenditure measurement, and includes a comprehensive review of what is already known about key methodological issues. The second section reviews the principal goals of collecting household-level expenditure data, outlining the various objectives that such surveys might satisfy, and implicitly or explicitly suggesting which goals are both feasible and important (especially in light of the existence of other data sources, like aggregate retail sales data, that might be able to answer some of the questions now addressed using household expenditure surveys).

The third section covers what is known about the existing Consumer Expenditure (CE) survey in the United States, with a focus on how well the survey tracks aggregate benchmarks, how it compares to similar surveys around the world, and how well it represents the underlying population being studied.

The fourth section reviews new modes of data collection, including the use of scanner data, Internet panels, and administrative data from government and private sources.

Coincident with the conference and the writing of this CRIW volume, the Bureau of Labor Statistics (BLS) sponsored a review by the National Academy of Sciences (NAS) of the CE redesign effort. That review panel began meeting around the same time that the CRIW conference was held, and released a detailed report in October 2012. The NAS panel members and staff had extensive interactions with authors of papers for this volume, and a number of the panelists and staff members attended the December 2011 conference at which preliminary versions of the papers were presented.

The NAS panel ultimately released a 260-page report on possible redesign alternatives that included numerous references to the work contained in this volume, and the panel requested (and received) permission to reproduce some of the exhibits prepared for the papers in this volume.[3] After reading

3. National Research Council (2013).

the panel's report, it seems clear that one conclusion upon which all panel members would agree is that a great deal of work remains to be done. Panel members were not able to agree fully on how best to proceed, and as a result the report contains a substantial dissent signed by a majority of the economists on the panel. (The panel included distinguished experts from a number of other fields including survey methodology, political science, and sociology, reflecting the broad scholarly uses to which the CE survey is put and the complexity it faces in achieving its goals.) Below we point out points of contact between the chapters in this volume and the NAS report.

In short, despite the important work undertaken by the panel, the question of how best to measure household-level expenditures remains unanswered, and this CRIW volume provides further evidence that while agreement may exist that fundamental redesign of household expenditure surveys is required, a great deal remains to be learned about what new methods of measurement would work better than those that have been employed in the past.

What Do We Already Know about Collecting Household Expenditure Data?

Chapter 1: "Asking Households about Expenditures: What Have We Learned?" (Thomas F. Crossley and Joachim K. Winter)

The starting point for the volume is a chapter by Thomas Crossley and Joachim Winter that summarizes what has been learned from previous studies about collecting household-level expenditure data. This extensive literature review is oriented around the key dimensions of the data collection process: survey mode, recall versus diary, disaggregation of expenditure categories, defining the response unit and choosing the respondent, reference period, the role of incentives, and the potential for reducing response errors in real time. This chapter's key contribution comes from its comprehensive approach and its global perspective; other chapters relating to data collection methodology per se generally make contributions on only one or two of these issues, and usually for a single country or a small number of countries.

Crossley and Winter are able to draw a number of conclusions about the various design decisions that have to be made in surveys that aim to collect household expenditure data. For example, they report evidence that diaries do not necessarily dominate recall surveys from a reporting perspective, and because there is incremental respondent burden in a diary, recall surveys may be preferable. They also find that research showing that higher levels of disaggregation improve recall may not be appropriate for the CE redesign question, because the CE already has much more detail than other surveys, and recent experiments with more aggregated categories finds aggregates that line up well with the more detailed CE (findings that are confirmed in chapters 13 and 14 by Michael Hurd and Susann Rohwedder later in this volume). Review of the literature on other questions about data collection

strategy yield more mixed results, and the authors identify several specific questions where more focused research is warranted.

Crossley and Winter's review describes the state of the international literature on expenditure surveys, as it stood at the time of our conference. Naturally, they describe research in a number of areas that figure prominently in the NAS report. For example, the NAS report suggests that the redesign of the CE must make use of incentives, and discusses the problem of respondents "learning to say no" (also called "motivated underreporting"), particularly when surveys have a cascading structure. Crossley and Winter review research on both these points.

The NAS report is specifically focused on the CE surveys, while the Crossley-Winter chapter is not. Consequently, the former highlights some CE issues that do not get much attention in Crossley-Winter. The most important of these is the sheer cognitive load of the CE interview survey. The CE is very detailed both in terms of the number of expenditure categories collected and in the follow-up information requested on purchases. The NAS report put a great deal of emphasis on the difficulty that respondents face in recalling the information requested by the interview survey. This is undoubtedly an important point, and a key reason why many suggest that the CE needs to be redesigned.

Goals for the Expenditure Survey Redesign

The CRIW conference in December 2011 contained a number of presentations illustrating the multiple goals of collecting household-level expenditure data. Four of those presentations are included as chapters here, providing a useful representation of goals from a number of different user perspectives. The first perspective is from the BLS itself, and aims to illuminate the original goal of the CE in generating weights for the construction of the Consumer Price Index (CPI). The chapter compares the CE against alternative approaches to generating expenditure weights. The other goals represent a range of academic applications, including studying household spending responses in a panel-data framework, using expenditures as an alternative to income when measuring inequality and poverty, and using expenditure data to model household-level spending responses to changes in prices and incomes.

Chapter 2: "Constructing a PCE-Weighted Consumer Price Index" (Caitlin Blair)

This chapter by Caitlin Blair seeks to answer the following question: How would our assessment of consumer price inflation change if we stopped using CE data to construct CPI expenditure weights and instead constructed weights using Bureau of Economic Analysis personal consumption expenditures (PCE)?

The question is important for CE redesign because of well-known divergence in rates of reporting across different types of spending in the CE.

For example, if the particular goods and services that are overweighted in the CE market basket are also the goods and services for which prices rose most rapidly, then the CPI will be biased upward relative to a PCE-weighted index.

Blair shows that the extent of CPI bias depends on the specific question being asked. If we adjust for conceptual and coverage differences between the two possible weighting schemes, then the results for overall inflation are not very different, at least for the time period being studied (2005–2010). If we do not adjust for conceptual differences, then some spending categories that are not well covered in the CE (especially employer-provided medical and spending on education) and that exhibit higher inflation over the study period do raise the overall inflation estimate by a noticeable amount—0.441 percentage points on the average twelve-month index change of 2.013 percentage points. This raises an important philosophical issue about what the CPI should be measuring—for example, do we want the (implicit) cost of employer-provided medical care to affect the CPI?

Chapter 3: "The Benefits of Panel Data in Consumer Expenditure Surveys" (Jonathan A. Parker, Nicholas S. Souleles, and Christopher D. Carroll)

The CE interview survey is unusual among national comprehensive household expenditure surveys in that it has a panel structure. Participating households are asked to complete five quarterly interviews. The first of these is designed primarily to bound recall; the subsequent four interviews are the basis for the data that is produced, yielding up to four observations on households spanning a period of up to a year. ("Up to" because many households do not complete all five interviews.) This chapter, by Jonathan Parker, Nicholas Souleles, and Christopher Carroll, assesses the value of this panel structure. They conclude that there is a strong case for retaining the panel element of the CE survey in any redesign, and that the panel structure of the CE interview survey is of value to both the core missions of the CE survey, such as price-index construction and poverty measurement, and to the research uses that the data serve.

The authors review the ways that the panel structure can improve measurement, for example, by reducing nonsampling error. One important aspect of this is that with a single recall period, surveys designers face a trade-off between greater recall error (with a longer recall period) or greater variability arising from purchase infrequency (with a shorter recall period). A design with repeated interviews on each sampled unit (a panel) relaxes this trade-off.

The authors also consider the role of the panel structure in the CE interview survey in supporting research. The key issues are heterogeneity and dynamics. The authors review how panel data allows for consistent estimation of parameters of interest in the presence of unobserved heterogeneity, and illustrate the argument with the example of studying the impact of stimulus tax rebates on spending. They also discuss how dynamic issues

such as habits in spending behavior and the degree of mobility in spending behavior can be studied with panel data on consumption.

The NAS report noted that the CPI does not utilize the panel nature of the current CE. On the other hand, the panel acknowledged that economic research and policy analysis was an important use of the CE and that the panel nature of the data was a key feature that makes the CE useful for such research. Each of the three prototype redesigns put forward in the report includes a panel component, although one of the options has data collection at just two points, and variable response periods (by expenditure category) at each point. This design may not produce data that is very useful for economic research and policy analysis, as the report acknowledges and the dissent to the main report further emphasizes.

Chapter 4: "The Evolution of Income, Consumption, and Leisure Inequality in the United States, 1980–2010" (Orazio Attanasio, Erik Hurst, and Luigi Pistaferri)

An influential set of papers culminating in Meyer and Sullivan (2012) has argued that, among poor households, income is badly mismeasured, while spending is less mismeasured; an obvious implication is that poverty researchers should use data on spending (e.g., from the CE survey) rather than on income to measure household well-being. Separately, a literature sparked by Krueger and Perri (2006) has shown that inequality in spending as measured by data from the interview component of the CE survey remained fairly stable over the past three decades in the United States, even as income inequality has widened dramatically; however, from its inception this literature has been plagued with doubts about whether its main result reflects increasing measurement error rather than true economic patterns.

This is the context for the chapter by Orazio Attanasio, Erik Hurst, and Luigi Pistaferri, who compare changes in US household spending inequality over the past thirty years to changes in measured income inequality over the same period, using data that they argue can (at least partly) overcome the criticisms that have been leveled at the CE data. Using an impressive variety of evidence, Attanasio, Hurst, and Pistaferri show that ever-increasing measurement error in the CE data explains the discrepancy between trends in spending inequality and income inequality. Specifically, they estimate spending inequality by (a) using a simple demand system that allows for measurement error; (b) using data from the diary component of the CE survey for items where past research has shown measurement error to be small; (c) using data on durables purchases, which also arguably have relatively small measurement error; and (d) using spending data from the Panel Study of Income Dynamics, which arguably are better measured than overall expenditures in the CE survey. With all four of these methods they find an increase in spending inequality that roughly matches the increase in income inequality, in sharp contrast to the pattern exhibited in the raw CE interview data. Together with the work of others whom they cite, this chapter provides a compelling illustration of

the importance of the growing measurement problems faced by expenditure surveys. The question (growing inequality in household well-being) is of great interest to policymakers and the public, but bad data has the potential to lead to profoundly mistaken conclusions about the nature, causes, and appropriate policy responses to the real economic changes that are taking place.

Chapter 5: "Using the CE to Model Household Demand"
(Laura Blow, Valérie Lechene, and Peter Levell)

The final chapter on CE goals is by Laura Blow, Valérie Lechene, and Peter Levell, and seeks to answer the following question: How does the availability of comprehensive household demographic and labor force data affect estimates of demand system parameters? The demand system parameters of interest are price and income elasticities, which are used extensively in structural policy models. These estimated elasticities are the key to predictions about general equilibrium effects of tax, transfer, and other government policies that affect consumer spending. The importance of this chapter for CE redesign is underscored by the fact that one could never properly estimate these elasticities without using household-level spending data, but one also needs demographic and labor force variables because the estimated demand parameters vary systematically based on those characteristics.

Blow, Lechene, and Levell estimate a number of different demand systems using a two-stage approach and different population subsamples. The commodities in their nondurable goods demand system are food in, food out, entertainment, apparel, utilities, and motor fuel. The authors conclude that the estimated demand system parameters are in fact dependent on the conditioning used to estimate the system, where conditioning refers to number of rooms in the housing unit, labor force participation, and stock of cars. The bottom line conclusion is that we need all the household-level data to be preserved in one place if we want to provide policymakers with appropriate demand system parameters for modeling policy changes.

The NAS panel emphasized the importance of nonexpenditure information collected in the CE, which is important for many types of research conducted using the CE; for example, the demand system estimation described in this chapter. Indeed, one key NAS panel recommendation involves better alignment of the timing for income and expenditure flows, which will improve the reliability of estimates that use income and other nonexpenditure information along with expenditure data.

Evaluating the Existing CE Survey

Much of the impetus for redesigning the CE survey comes from a growing realization that the current BLS methodology leaves much to be desired in representing aggregate household spending. Assessing the extent to which the CE diverges from aggregate benchmarks requires a comprehensive reconciliation of exactly what is being measured, and a comparison of how dif-

ferent approaches using the CE itself (diary versus interview) give different answers. Both the fact that CE aggregates are below aggregate benchmarks and the fact that the discrepancies are worsening has motivated further investigations into whether the same phenomenon is occurring in other similar surveys around the world, and to what extent underrepresentation of the wealthiest families may be affecting comparisons against aggregate totals.

Chapter 6: "Understanding the Relationship: CE Survey and PCE" (William Passero, Thesia I. Garner, and Clinton McCully)

The chapter by William Passero, Thesia I. Garner, and Clinton McCully seeks to answer the following question: How does the new concordance between CE and personal consumption expenditures (PCE) developed by BLS and Bureau of Economic Analysis (BEA) staff affect how well the two data series track each other over time? The authors of this chapter have written extensively about CE versus PCE aggregates in previous papers, and they focus this chapter on the specific issue of how that concordance is affected by the new BEA spending categories introduced a few years ago. The importance of this chapter for CE redesign is paramount, because assessing whether the CE is comprehensively capturing household spending necessarily begins with comparing aggregates across spending categories and time.

Passero, Garner, and McCully focus on two aspects of the question. First, how much conceptual overlap is there between CE and PCE? Second, how do the ratios of comparable CE to PCE aggregates vary across spending categories and time periods? The conceptual differences between the two data sets are significant. As of 2010, only 62 percent of PCE expenditures will, in principle, be captured by the CE, and only 80 percent of CE expenditures will, in principle, show up in the PCE. These comparability ratios are highest (94 percent for both PCE and CE) for nondurable goods and lowest (48 percent for PCE, 73 percent for the CE) for spending on services. Regarding trends over time and focusing on comparable goods and services only, the authors conclude that CE to PCE ratios have steadily decreased. For total comparable goods and services, CE to PCE ratios decreased from 84 percent in 1992 to 74 percent in 2010. The greatest decline in CE to PCE ratios is for durables, with a decrease of 24 percentage points. Ratios for comparable services dropped the least, with a percentage decrease of 10 percentage points. The NAS panel requested (and were granted) the ability to cite numbers from this chapter in their report, as part of the core evidence about deterioration of CE representativeness over time.

Chapter 7: "The Validity of Consumption Data: Are the Consumer Expenditure Interview and Diary Surveys Informative?" (Adam Bee, Bruce D. Meyer, and James X. Sullivan)

This chapter, by Adam Bee, Bruce D. Meyer, and James X. Sullivan, provides an assessment of the quality of the data collected by the current CE

surveys. While data generated by the CE surveys has been assessed against various benchmarks before, the key contribution of this chapter is to assess the CE interview survey and CE diary survey separately (past analyses have often assessed a combination of the two). This approach delivers a number of insights. The most of important is that in careful comparisons to the national accounts, the interview survey appears to perform better than the diary survey. Many large categories of expenditure seem to be well measured in the interview survey, in that the ratio of implied aggregate spending to the relevant national accounts figures is close to 1 and stable over time. The authors note that the diary data also contain many more reports of zero expenditure in a consumption category. These zeros, which may be related to purchase infrequency, cause significant problems when using the data to assess levels of poverty and inequality. Overall, the authors argue that for many purposes the interview data may be superior to the diary data.

In additional analysis, the authors show that the CE compares well to external sources on ownership and value of durables, particularly homes and cars. This is important for analysis that requires an imputation of households' total consumption, including service flows from durables. Such a measure is required, for example, in assessing living standards and poverty. They also provide some evidence that the CE interview survey sample is representative of the target population along many dimensions, although they acknowledge concerns about underrepresentation at the top of the income distribution, which are raised in the next chapter.

The main NAS report rejects the central conclusion of this chapter—that by many measures the current CE interview survey data are superior to the data from the diary survey. The report argues that it is not possible to determine which mode is inherently better, and all of the prototype redesigns developed by the panel include a significant diary (or journal) component. Those proposals do, however, include significant changes to the current diary mode of the CE, including the adoption of technologies for self-administrated data collection (including tablet computers and home scanners).

Overall, the NAS report calls for a greater, rather than lesser, role for diary modes of data collection. The dissent to the main report, which was cowritten by one of the authors of this chapter, expresses a reservation about a move to greater reliance on diary-based data collection. The dissent points to the evidence in this chapter, and to earlier evidence on the relative quality of diary and recall methods summarized in chapter 1 by Crossley and Winter.

Chapter 8: "Is the Consumer Expenditure Survey Representative by Income?" (John Sabelhaus, David Johnson, Stephen Ash, David Swanson, Thesia I. Garner, John Greenlees, and Steve Henderson)

The underreporting of expenditures was cited as a major motivation for the NAS review of the CE redesign effort. This chapter, by John Sabelhaus, David Johnson, Stephen Ash, David Swanson, Thesia I. Garner, John

Greenlees, and Steve Henderson begins with the observation that underreporting can arise in two main ways. It could be that high-income, and hence high-spending, households are underrepresented in the CE sample, or it could be that some or all households underreport their spending. Of course, both sources of error could be operative.

The authors bring a valuable new data source to bear on the question of the importance of these two sources of error. This data set links sampled units from the CE interview survey, both those that responded and those that did not, to their zip-code level average adjusted gross income (AGI). This allows the authors to examine directly response rates by AGI percentile income groups. It turns out that the CE response rate is fairly constant between the 10th and 90th percentile of AGI, but that there is less nonresponse in the bottom decile and significantly more nonresponse above the 90th percentile. Households in the top 5 percent of zip code-mean AGI are about 10 percentage points less likely to respond to the survey. This is the first direct evidence that high-income households are underrepresented in the CE sample.

Nevertheless, the authors argue that the underrepresentation of high-income households in the CE sample cannot close all the gap between national accounts expenditure totals and aggregates derived from CE data: multiplying the missing income by estimates of the marginal propensity to spend for the high-income group does not deliver enough extra spending. Thus it seems that underreporting of spending also plays a role. The authors note that, given income, spending reports of the lowest-income households in the CE survey are implausibly high, and the spending reports of the highest-income households are implausibly low (implying rates of wealth accumulation that are not consistent with wealth surveys).

Thus it seems that both underrepresentation of high-income households and underreporting of spending by high-income households contribute to overall underreporting of spending in the CE survey. The authors conclude that the CE design effort must consider strategies for addressing these twin problems and discuss several, including the oversampling of more affluent households (as in the Survey of Consumer Finances) and the streamlining of the data collection process to make it feasible for high-income households to accurately estimate their spending. The main NAS report and the accompanying dissent both raised the issue of oversampling high-income families, because BLS did not emphasize the importance of that component in the redesign proposals that were given to the panel. The second issue, streamlining data collection, is also a key theme in the NAS recommendations.

Chapter 9: "A Comparison of Micro and Macro Expenditure Measures across Countries Using Differing Survey Methods" (Garry Barrett, Peter Levell, and Kevin Milligan)

A final empirical perspective on the current CE is the chapter by Garry Barrett, Peter Levell, and Kevin Milligan. These authors analyze how differ-

ences in expenditure data collection methodologies across countries are reflected in differences in the quality of data collected. The measures of data quality that the authors consider include both response rates (fraction of selected respondents who participate in the survey) and coverage rates (ratios of survey spending aggregates to published national account aggregates for the same categories of spending). The coverage rates estimates for the CE are consistent with findings in other chapters in this volume, and also permeate the NAS report. The importance of this chapter for CE redesign is that we may be able to learn something from divergent experiences across countries. Barrett, Levell, and Milligan choose four Anglophone countries for their comparison: Australia, Canada, the United Kingdom, and the United States. Their chapter begins with a concise description of how the four surveys differ, both in terms of how the samples are drawn and how the surveys are conducted. The authors show a general deterioration in survey response rates across all four countries since the 1980s, but a general decline in coverage only for the United Kingdom and the United States. That is, the ratio of survey to aggregate spending in Australia and Canada has not deteriorated over time. One interesting possibility the authors consider is that the decline in coverage rates may be correlated with shifting income distributions. If households in the top 1 percent of the income distribution are less likely to participate in the survey, then an increasing share of income going to the top 1 percent will cause a small drop in response rates but a large drop in coverage. The authors find some evidence that this helps explain differences in coverage trends across countries.

Alternative Approaches to Data Collection

The CRIW conference and this volume were motivated by the prospects of addressing the shortcomings of current data collection methodologies, and at the same time improving the ability to achieve the agreed upon goals for collecting the data in the first place. Simultaneously improving measurement and achieving multiple goals (while still adhering to a statistical agency's budget constraint) will require considering new approaches to collecting data, which means moving beyond the traditional survey setting. Toward that end, the remaining seven chapters in this volume are focused on methodological changes such as real-time cash-flow reconciliation (balance-edit) to help minimize misreporting, combining survey and administrative data, self-interviews using the Internet, the effect of allowing respondents to choose reporting periods, and scanner technologies. Many of these possibilities also received attention in the NAS report, and the report encouraged the testing of new technologies as means of improving particularly self-completed data collection methodologies.

Yet another new way of measuring a household's total spending has emerged from Scandinavian countries in which government agencies col-

lect extensive information about each taxpayer in a centralized database. In principle, if perfect data on wealth and income data over time were available to tax authorities, it would be possible (for example) to compute the amount of an individual's spending by presuming that any non-capital-gains-related increase in wealth reflected a choice to spend less than measured after-tax income (the residual method). Of course, there are many complexities in implementing the residual method in practice, ranging from the difficulty of observing capital gains and losses to the existence of forms of income and wealth that are not reported to the tax authorities.

Both Denmark and Sweden happen to have conducted traditional consumer expenditure surveys during the period when the national registry data are available. And in both cases, scholars contributing to this volume have managed to link the data for participants in those expenditure surveys to the national registry data for the surveyed individuals. These two chapters differ somewhat from the others in this section; while the method is indeed new, it is not one that is likely to be implementable (or at least not very quickly implementable) in countries that have not built national registry systems. For this reason, and because this method does provide the detailed information on expenditure by category that is required for many uses of the CE, this approach did not get much attention in the NAS report. Nevertheless, these chapters are also unique in that they provide the only method we know of for testing the "external validity" of existing survey methods. For this reason, they provide a useful background for the other chapters in this section, so we begin with them.

Chapter 10: "Measuring the Accuracy of Survey Responses Using Administrative Register Data: Evidence from Denmark" (Claus Thustrup Kreiner, David Dreyer Lassen, and Søren Leth-Petersen)

For Denmark, the chapter by Claus Thustrup Kreiner, David Dreyer Lassen, and Søren Leth-Petersen reports an extensive set of comparisons between the registry-based "residual" method of measuring spending and the survey-based method, with the explicit aim of extracting lessons about the pitfalls of surveys. On the whole, they find a disturbingly small correlation between spending as measured using the residual method and spending as measured by the survey; according to one metric, a regression of registry-measured spending on survey-measured spending yields a coefficient of 0.791 with an R^2 of only 0.46.

Among the many other interesting results in this chapter, one stands out as possibly the most important: answers to the expenditure survey's question about the household's total income had remarkably little correlation with income as measured by the tax authorities. The authors make a persuasive case that the data from the tax records are likely to be fairly accurate. This result is disturbing because almost all existing expenditure surveys rely on self-reported measures of income (like the one in the Danish survey) for a

host of benchmarking and other purposes. Furthermore, total household income is much easier to compute than many of the other items about which households are questioned on such surveys. If households cannot accurately answer even a (comparatively) simple question like what their income was for the prior year, it is difficult to have confidence that the answers they are giving to other questions are accurate.

The authors examine whether various plausible kinds of confusion (between gross and net income, for example) might explain their disturbing results, but in the end they are not able to resolve the problem. They also show that the errors are nonclassical (that is, they are correlated in ex ante unknowable ways with characteristics of the population), which presents thorny statistical problems in figuring out appropriate methods of correcting for error.

The authors point out that the Danish government has encouraged the use of these data for research purposes, and a growing number of academic studies and statistical analyses have been conducted using them. For researchers who bring appropriate funding to the table, and who can make contact with a collaborator who can gain access to the data (naturally, access to the data is tightly restricted for security reasons), Denmark could become a uniquely useful "laboratory" for conducting experiments on what works and what does not for survey measurement. For example, one question that the Committee on National Statistics panel report highlighted as crucially important, but despaired of as nearly unknowable, was the dynamic properties of survey-response error. That is, if a person makes an error of a given size in a given survey, if that person is reinterviewed at some later date are they likely to make exactly the same error, or an independent error, or something else? As the authors point out, questions of this type could be investigated by commissioning a study using Danish data, where "truth" is known to a reasonable degree of accuracy.

Chapter 11: "Judging the Quality of Survey Data by Comparison with 'Truth' as Measured by Administrative Records: Evidence from Sweden" (Ralph Koijen, Stijn Van Nieuwerburgh, and Roine Vestman)

The chapter by Ralph Koijen, StijnVan Nieuwerburgh, and Roine Vestman takes up the case of Sweden. In principle, the data available to the Swedish government are even more impressive than in Denmark; this is a legacy of the Swedish wealth tax (which was abolished in 2007). In order to implement such a tax the authorities needed to be able to compute the net worth of each individual. For assessing individual tax obligations, an automatic reporting procedure from financial institutions to the tax authorities was set up, resulting in a mechanism by which highly disaggregated information on the income and wealth of all households flowed to government records. Individual financial asset, mutual fund, and real estate portfolios are provided at the single property and security level during the period covered by the expenditure survey.

Since spending (in this approach) is measured by comparing income to the change in wealth, being able to determine the extent to which wealth has changed as a result of capital gains or losses (and not a result of active saving or dissaving) is a crucial advantage. Other studies (including the Danish registry study) have had to make assumptions about the size of capital gains and losses, typically assuming that a fixed aggregate rate of return applied to all assets of a particular class. (See, e.g., Maki and Palumbo [2001], and chapter 14 of this volume by Hurd and Rohwedder).

The authors find that properly accounting for the idiosyncratic capital gains and losses does make a substantial difference to measured expenditures for many households, and that (intuitively) this problem is larger the greater a household's wealth.

Overall, they find that the mean and median levels of spending are similar in the two sources (their registry computations and the survey). Again, however, at the level of individual households the results are disturbing. Even among the subgroup that the authors identify as likely the best measured in their data (renters measured in December), the correlation between survey-based and registry-based consumption is only about 0.5, and the correlation is substantially lower for other groups of households. Indeed, and somewhat surprisingly, the relation between their registry-based measure of spending and the survey-based measure at the level of individual households is looser than the corresponding relation in the Danish study. This is true even though the authors present evidence that the Swedish registry's information on capital gains and losses does improve the coherence between the Swedish registry-based measure of spending and the survey-based measure. A possible interpretation is that while the Swedish registry-based data is better, the Swedish survey-based data is worse than in Denmark. Or perhaps some other aspect of the Swedish registry data is worse.

One hint that the Swedish survey data may be seriously problematic is that, among persons who are known (from the reliable national registry records) to have purchased a vehicle during the last twelve months before the date of the survey, only 71.2 percent of survey respondents report having purchased this vehicle. Since vehicle purchases have long been viewed as one of the most reliable kinds of data obtained by household surveys, this is surprising, and suggests either that the Swedish survey was unusually inaccurate or that the presumption among researchers that vehicle purchases are measured well is misplaced.

One particular finding resonates with the message of Aguiar and Bils (2011): the authors find that, in the survey, spending is particularly understated for richer households. It is not obvious a priori that the biases in a Swedish spending survey should be similar to those in an American survey, and this result suggests that it is not unreasonable to hope more broadly that lessons obtained in one country may apply to other countries as well.

Chapter 12: "Exploring a Balance Edit Approach in the Consumer Expenditure Quarterly Interview Survey" (Scott Fricker, Brandon Kopp, and Nhien To)

Reporting detailed spending is a difficult task for households, and so it is perhaps unsurprising that some, or perhaps many, households underreport their spending. Some comprehensive household spending surveys include a "balance edit" as a data-control measure. A balance edit compares a household's reports of spending, income, and changes in assets and liabilities. These totals are, of course, linked by the household's budget constraint: the difference between income and spending must be flows to or from assets and liabilities. Where the reported elements of a household's budget constraint are out of balance by a predetermined amount, respondents are given the opportunity to review and revise their responses. Early versions of the CE survey had such a measure, but it was eliminated in the major redesign of 1972, in part because it was thought to be infeasible to conduct the balance edit in the context of the quarterly interview survey introduced at that time. However, research based on other surveys suggests that a balance edit can be useful in improving households' reports of spending and income.

This chapter, by Scott Fricker, Brandon Koop, and Nhien To, reports on a small-scale test of a modified version of the CE interview survey with a balance edit procedure. The test was conducted in the Office of Survey Methods Research Laboratory, and this allowed the authors to use cognitive testing methods and participant debriefing to investigate not only if the balance edit works, but how it works.

In the experiment, the balance edit improved the balance for a majority of participants, but only a small fraction of respondents were able to achieve balance. Debriefing revealed very heterogeneous comprehension of, and reaction to, the balance edit. While most respondents understood the measure and had neutral or positive reactions to it, there was a group of respondents who struggled to understand the balance edit and a second group who had a negative reaction to it. The latter included individuals whose spending exceeded their income. The authors conclude that balance edit procedures have some potential for improving data quality, but that there are significant issues to be considered in the design and implementation of any such procedure and the usefulness of the procedure is likely to depend on specific details of a redesigned CE survey.

The use of a balance edit or similar methods to improve data quality did not get much attention in the NAS report. The report does note that this method has recently been dropped from the Canadian Budget Survey, as it transited to greater reliance on diaries. The dissent to the main report felt that the report could have put greater emphasis on ways to monitor data quality, and cited the use of budget balance as one possible approach.

Chapter 13: "Measuring Total Household Spending in a Monthly Internet Survey: Evidence from the American Life Panel" (Michael D. Hurd and Susann Rohwedder)

The first of two chapters by Michael Hurd and Susann Rohwedder presents a potentially revolutionary new measurement tool for household expenditures: an Internet panel. Panel participants agree to answer questions using an Internet-enabled device (they are given such a device if they do not have one) on a regular schedule in exchange for a payment to compensate them for their time and effort. While it seems reasonable to worry about the representativeness of such a sample, at some point, as more and more daily routines of life get integrated into the Internet, it may become more reasonable to question the representativeness of a sample not conducted using Internet tools. (This point is especially compelling given the plummeting response rates for non-Internet-based survey methods.) The proliferation of Internet-based collection methods for such data is creating the knowledge needed to adjust the sample to correct for bias. A proof of the effectiveness of such sample adjustment came from the 2012 elections in the United States: a prominent expert ranked the entirely Internet-based Google Consumer Polls as the second-most accurate among all pollsters using all survey methods.[4]

Hurd and Rohwedder report a host of interesting results obtained by adding a carefully considered set of spending questions (based on experience gained from the Health and Retirement Study) to the financial crisis surveys that they began conducting in the American Life Panel (which interviews about 2,500 households on a regular basis) immediately after the onset of the recent financial crisis. Using a variety of methodological innovations, they produce a measure that appears to capture the bulk of the spending measured by the far more expensive and elaborate CE survey. Furthermore, because of the panel structure of their survey, they can observe changes in spending patterns in response to economic events like movements in the stock market.

In their first financial crisis survey (November 2008), 73 percent of households reported that they had reduced spending because of the economic crisis. Prompted by this striking result, and by their knowledge that understanding the spending response to the crisis would be critical for analyzing it, they began working to establish a monthly interview schedule for spending questions, which was implemented in May 2009, with monthly data available thereafter. A particularly interesting finding is the discrepancy between the recovery in spending at the median and at the mean. They find that both mean and median spending reached a trough in May 2010, but that (by the time the data sample used in their chapter ended) median monthly spend-

4. Silver (2012).

ing had recovered only 8 percent from its trough while mean spending had recovered by 11 percent. These are the kinds of high-frequency results that heretofore have been possible to calculate only years later when (for example) the cleaned and edited CE survey becomes available. Their chapter shows the potential for getting at least a rough-and-ready measure of how distributions are changing nearly in real time.

Chapter 14: "Wealth Dynamics and Active Saving at Older Ages" (Michael Hurd and Susann Rohwedder)

A second contribution by Michael Hurd and Susan Rohwedder explores a classic question in the economics of life cycle behavior—Do most people aim to spend their wealth before they die?—using another relatively new tool for measuring spending. Over the past decade, the US Health and Retirement Study (HRS) has added a battery of spending and other questions (the Consumption and Activities Mail Survey [CAMS]) to its core household questionnaire. Hurd and Rohwedder show that the HRS's CAMS data match the spending of similarly aged households in the CE survey reasonably well (especially given the vastly smaller resources employed in the CAMS measurement exercise), with the CAMS measure generally exceeding the corresponding CE measure by between 8 and 16 percentage points. (Since a primary problem of the CE survey is that it misses substantial amounts of spending [cf. Attanasio, Hurst, and Pistaferri in chapter 4, this volume, and the papers cited therein], it is even possible that the CAMS comes closer to the truth than the CE does.)

Turning to the motivating question (do people draw down their wealth as they age), the chapter is able to use the CAMS measure of spending in combination with the HRS's fairly complete measures of income to construct a measure of "active saving" (the difference between income and expenditures). The authors then compare that measure to the results obtained by examining the changes in wealth across survey waves. They find broadly consistent results: while single individuals do appear to be drawing down their wealth, elderly couples continue to save (presumably in order to finance the spending of the survivor when one of them dies).

The chapter illustrates the point that adding carefully considered spending questions to existing surveys may not be as costly as once thought, and that important topics can be studied using such questions. The interesting contrast is between the "bottom up" survey method traditionally employed by CE surveys (asking about spending category-by-category for narrowly defined categories of products), and the more aggregated approach in CAMS-type surveys, which aims at a "big picture" and does not worry about getting spending details. While results from big picture questions may not be useful in constructing basket weights for price indices, the answers to such questions are key for understanding issues of saving, overall inequality, and household finances.

Chapter 15: "Measuring Household Spending and Payment Habits: The Role of 'Typical' and 'Specific' Time Frames in Survey Questions" (Marco Angrisani, Arie Kapteyn, and Scott Schuh)

In designing recall expenditure questions, two important issues are the length of the recall period (a week? a month?) and whether the question should refer to a specific period (such as last week) or a "typical" or "usual" period. Survey response theory tells us that different question designs may induce very different response styles. Longer recall periods and typical periods are more likely to lead to rate-based estimation, while respondents are more likely to enumerate when faced with shorter and specific recall periods. Short periods suffer from less recall error, but exhibit higher variability due to purchase infrequency. Specific recall periods may exhibit variability due to purchase infrequency or seasonal effects. How different designs perform is ultimately an empirical question.

This chapter, by Marco Angrisani, Arie Kapteyn, and Scott Schuh, reports on an experimental module in the American Life Panel (ALP). Respondents were asked the number and amount of purchases by different payment methods (debit cards, cash, credit card, and personal check). Respondents were interviewed four times. For each respondent, subsequent interviews switched between typical and specific formats, with the format of the initial interview randomly assigned. Within each interview respondents were asked about different recall periods (a day, a week, a month, and a year), with the order of different periods randomly assigned. Results from the first round of interviews are reported in this chapter.

On average, respondents report higher numbers of payments and greater amounts for short recall periods (a day or a week). For most payment methods, the probability of reporting nonzero payments is higher for typical than for specific periods, but amounts spent are systematically lower for typical periods. These results illustrate the important influence of recall period type and length on reporting behavior.

This chapter shows that type (specific/typical) and length of recall period greatly affect household-reporting behavior. The current CE interview survey uses a three-month recall period for most goods and the NAS report argues that this is very long for actual recall of many items. One of the prototype redesigns moves away from a common reporting period for all expenditure categories. On the other hand, the issue of specific versus typical periods does not seem to have received much attention in the report.

Chapter 16: "The Potential Use of In-Home Scanner Technology for Budget Surveys" (Andrew Leicester)

Another novel mode of data collection is the use of in-home scanners to record information in individual purchases; market research firms have

developed these devices as a tool for measuring the effects of advertising and for other commercial purposes. This chapter by Andrew Leicester considers how scanner data might be used in the context of a comprehensive survey of household expenditures.

His chapter yields a number of insights that could guide future choices by statistical agencies. One disappointing result is that spending patterns of different households within the same store are quite different. This is discouraging because if all consumers had the same spending patterns for a given store, then it would be possible to impute to a household detailed spending patterns by category of goods based just on the distribution of their spending across store types. Leicester's result shows that this would lead to mistakes (at least at the level of an individual household).

Leicester also finds results that could be helpful in understanding differences between survey results from interview surveys (which typically cover an extended time period, like three months) and results from diary surveys (which typically cover a shorter period, like two weeks). For example, over any given two-week period Leicester finds that a high proportion of households buy no fish. If household-specific expenditure weights for a CPI were constructed using such data (as, Leicester reports, has been done), the price of fish would have no effect on the computed household-specific inflation rate for these households. Yet, Leicester shows that when the time frame is extended (at its longest, to a year), the proportion of households who buy no fish is much lower. Broadly speaking, Leicester's results tend to suggest that in order to provide a reasonably accurate measure of a household's "true" spending patterns (for purposes like constructing individual- or group-specific CPIs), it will be necessary to collect data over an extended time interval, perhaps as long as a year. Two-week diary surveys are not adequate for this purpose.

This is an important conclusion, in part because it speaks directly to a major source of dissent among members of the Committee on National Statistics panel that BLS convened to provide advice on revising the CE survey. The dissenting members believed that diary survey approaches should be abandoned because even if the data obtained from them were accurate, the time frame covered by diary surveys is too short for the data to have any meaningful economic use. Leicester's results bolster the dissenters' argument by showing that the expenditures that a household makes over a two-week period are very far from being a good picture of their expenditure patterns over an entire year. Indeed, he shows that patterns of expenditures are markedly different even between the quarterly and the annual frequency. This suggests that to obtain a reasonably useful picture of a household's expenditure patterns it may be necessary to collect data for a period as long as a full year.

References

Aguiar, Mark A., and Mark Bils. 2011. "Has Consumption Inequality Mirrored Income Inequality?" NBER Working Paper no. 16807, Cambridge, MA.

Krueger, Dirk, and Fabrizio Perri. 2006. "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory." *Review of Economic Studies* 73 (1): 163–93.

Maki, Dean M., and Michael G. Palumbo. 2001. "Disentangling the Wealth Effect: A Cohort Analysis of Household Saving in the 1990s." Finance and Economics Discussion Series 2001–21, Board of Governors of the Federal Reserve System.

McCall, Robert. 2012. "Development of International Guidelines and Frameworks for Micro Statistics on Household Income, Consumption and Wealth." Presentation at 32nd General Conference of the International Association for Research in Income and Wealth. http://www.iariw.org/ papers/2012/McCollPaper.pdf.

Meyer, Bruce D., and James X. Sullivan. 2012. "Winning the War: Poverty from the Great Society to the Great Recession." *Brookings Papers on Economic Activity* 45 (2): 133–200.

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

Silver, Nate. 2012. "Which Polls Fared Best and Worst in the 2012 Presidential Race?" http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/.

Stiglitz, J. E., A. Sen, and J. P. Fitoussi. 2009. "Report by the Commission on the Measurement of Economic Performance and Social Progress." Council on Foreign Relations. http://www.cfr.org/world/report-commission-measurement-economic-performance-social-progress/p22847.