Asking Households About Expenditures: What Have We Learned?

Thomas F. Crossley

Koc University
Institute for Fiscal Studies
University of Cambridge

Joachim K. Winter

University of Munich

This draft: April 2012

Abstract: When designing household surveys, including surveys that measure consumption expenditure, numerous choices need to be made. Which survey mode should be used? Do recall questions or diaries provide more reliable expenditure data? How should the concept of a household be defined? How should the length of the recall period, the level of aggregation of expenditure items, and the response format be chosen? How are responses affected by incentives? Can computer-assisted surveys be used to reduce or correct response error in real time? In this paper, we provide a selective review of the literature on these questions. We also suggest some promising directions for future research.

Keywords: expenditure, consumption, measurement error, survey data

JEL classification: C81, D12

Acknowledgements: This paper was written for the Conference on Improving the Measurement of Consumer Expenditures sponsored by Conference on Research in Income and Wealth and the National Bureau of Economic Research, with support from the Centre for Microdata Methods and Practice. We thank conference participants for many helpful comments. Some of this work on this paper was co-funded by the ESRC-funded Centre for Microeconomic Analysis of Public Policy (CPP, reference RES-544-28-5001).

1. Introduction

The importance of expenditure data to a wide range of important areas of both basic research and policy analysis has been well argued elsewhere (for example, Deaton and Grosh, 2000; Browning, Crossley and Weber, 2003). We believe the case is broadly accepted. At the same time, there is mounting evidence of problems with the household budget surveys conducted by national statistical agencies in many countries. The U.S. Consumer Expenditure Survey exhibits declining response rates and a diminishing correspondence to national account aggregates, and similar patterns have emerged in the budget surveys of other nations (see the evidence in Barrett, Levell and Milligan, 2012, this volume). There is also substantial evidence that survey design and data quality affect substantive conclusions about important research questions. A good example is the study of the evolution of inequality in the U.S. by Attanasio, Battistin and Ichimura (2004).

These facts have led to a number of initiatives that investigate what might be done to improve the quality of expenditure data collected and available for research and other purposes. These initiatives include the NBER-CRIW conference to which this paper is a contribution. Another is the Gemini Project, a multi-year, interdisciplinary research effort initiated by the U.S. Bureau of Labor Statistics in 2009 to inform the redesign of the Consumer Expenditure (CE) surveys. The aim of the CE survey redesign is to improve data quality through a verifiable reduction in measurement error – particularly error caused by underreporting. Papers written for the Gemini Project, or presented at its regular meetings, investigate many of the issues covered in this survey.[1]

In fact, researchers and survey designers have been studying alternative ways of collecting household expenditure data for many years. The resulting literature is very disperse, distributed over many years, many countries and multiple academic disciplines. Given the renewed attention that the

[1] Details on the Gemini Project, including project papers, other materials, and recommendations, can be found online at http://www.bls.gov/cex/geminimaterials.htm (last accessed March 8, 2012).

collection of expenditure data is now receiving, it seems timely to try to bring that literature together in an accessible way. This paper is an attempt to do so.

Like any short survey, this paper is necessarily selective and circumscribed. It is aimed primarily at economists and researchers that traditionally analyze expenditure data, but who are becoming increasingly involved in the design of data collection. There is experience with the collection of expenditure data in both developed countries and in developing countries. Some of the issues are common and others specific; our focus is tilted towards evidence from developed countries, but we mention evidence from less developed countries when it seems to us particularly useful. Deaton and Grosh (2000) discuss many more results from developing countries.

2. The Design of Expenditure Surveys: The Evidence

In this section, we review existing studies on various design choices that arise in expenditure surveys: Survey mode; recall vs. diary surveys; response formats for recall questions; surveys that predict aggregates from components; the level of aggregation of expenditure items; the definition of the response unit; the reference period; the role of incentives; and approaches to reduce or correct response errors in real time.

2.1 Survey mode

A first important decision in the administration of household surveys concerns the survey mode, the most common options being personal (face-to-face) interviews, telephone interviews, and self-administered questionnaires. All three modes could be based on a paper questionnaire or a computer interface. Other than for self-administered surveys, including 'leave behind' questionnaires in personal interviews, the use of paper questionnaires has become rare. Self-administered questionnaires are increasingly administered using the internet. There is a large literature on how the survey mode affects responses which we certainly cannot review here; see Tourangeau et al. (2000) for an overview.[2]

Key aspects of the interaction between survey mode and response behavior that are relevant for expenditure measurement concern the comprehension of survey questions (since an interviewer can provide clarification of difficult questions should this be allowed by the survey protocol) and the sensitivity or confidentiality of the target quantities (since the presence of an interviewer might increase such concerns). A third consideration is that self-administered surveys make it easier for respondents to look up information on hard-to-recall quantities such as asset holdings should they be willing to do so. While there is a large literature on mode effects in survey research, there is little systematic evidence when it comes to asking for consumption expenditure.

Models of survey response behavior suggest that written surveys enhance respondents' understanding of survey questions relative to oral presentations. Kemsley (1965) noted that expenditure data collected by self-completed diaries does not exhibit statistically significant interviewer effects while data collected by recall interviews does (the interviewer still plays a role with the diaries, in that they drop off, explain, review and collect them). While he interpreted this as sign of lower quality (e.g. greater subjectivity) in the latter, there is also the possibility that interviewer presence might have a positive (though uneven) effect on respondent comprehension, or on other steps in the response process.

Confidentiality concerns will be more relevant in personal than in self-administered interviews. This would suggest that personal interviews should result in lower estimates on potentially sensitive goods, such as alcohol, and there is some evidence that this is the case (Silberstein and Scott, 1991). On the other hand, a consistent finding of many studies is that response rates to total household expenditure questions are higher than response rates to comparable income questions

[2] Several chapters in this volume deal with using the internet to elicit consumption expenditure, so in this chapter we do not discuss issues that are specific to internet surveys in detail.

(Browning, Crossley, and Weber, 2003) suggesting that respondents view questions about broad categories of expenditure as being less sensitive than comparable income questions. This interpretation has been corroborated in the recent U.K. focus group studies summarized by d'Ardenne and Blake (2011).

Essig and Winter (2009) conducted a controlled survey experiment on mode effects in household surveys. In the data from the German SAVE household survey they analyzed, a random group of respondents answered sensitive questions, including those on household income and assets, using a questionnaire that was left behind by the interviewer rather than as part of the main interview so that it could be answered in private and independently of the rest of the survey interview. In comparison to the computer-assisted personal interview (CAPI) mode, rates of non-response were lower in the paper-and-pencil drop-off questionnaire. This effect was pronounced for all six asset categories they analysed, while there was no significant effect on item nonresponse to the question on household net income. This result suggests that the strength of mode effects is not constant across different target quantities that vary in sensitivity. An alternative interpretation is that households were willing to look up their asset holdings in their records when the leave-behind questionnaire allowed them to do so, the premise being that asset holdings are more difficult to recall from memory during a survey interview than income.

Bonke and Fallesen (2010) offered survey respondents the choice between answering a telephone and an internet survey (their main research interest was the role of incentives; see below). The study included both expenditure and time use questions. Overall, they found that response quality was higher when respondents chose to participate in the internet survey over the telephone interview. Due to the self-selection of respondents, the mode effect cannot however be interpreted causally.

Safir and Goldenberg (2011) analyze variation in the mode of administration of the Consumer Expenditure Survey. They argue while telephone interviewing may impact the quality of CE data relative to that obtained by personal visit interviewing, mode effects can be mitigated by using 'recall aids' in both modes, for example, through a mailed information booklet and user-friendly checklists for records and receipts.

An important concern with using a mixed-mode design is that response rates might be different across modes. Shin, Johnson, and Rao (2011) compare unit and item response rates in web and mail survey modes in the 2008 Gallup Health Panel Survey. They find that the web survey mode produces a lower unit response rate compared to the mail mode. However, the web mode elicits higher data quality in terms of item response to both closed and open-ended questions. These mode effects on data quality remain after controlling for socio-demographic characteristics.

Summary. The survey mode influences response behavior via various channels, the most important being comprehension of the questionnaire; ease of recall and information look-up; and confidentiality and sensitivity of the responses. Given that these channels interact, there cannot in our view be an easy answer to the question of which survey mode works best when it comes to consumption expenditure. Moreover, while there is a large literature on mode effects in survey research, we are not aware of a systematic, controlled experimental study of how survey mode affects response quality in expenditure surveys along these channels.

2.2 Strategies for the collection of expenditure data: recall vs. diary

A second fundamental design choice in expenditure measurement is whether respondents are asked to report how much they spend on consumption goods in a certain period (the recall approach) or whether they fill in a diary over a certain period of time in which they record every single expense (the diary approach). A final strategy for measuring consumer expenditure is the use of home scanner data; this approach is covered in detail by Leicester (2012, this volume) and therefore not discussed here. Recall surveys of expenditure have typically been conducted by interviewers, raising the mode issues discussed in the last section. However, there have been a number of recent experiments with recall expenditure surveys administered by mail and by internet (Hurd and Rohwedder 2009, 2010) so that collection mode and collection method are no longer tightly linked. For each of the recall and diary methods, there are additional design choices to be made, such as the length of the reference period (both recall and diary) and the level of disaggregation and the response format. We will review these aspects in the following sections; but first, we review evidence that concerns the choice between recall and diary approaches.

Problems with recall methods

The literature on survey response behavior noted early on that questions that require recalling quantities from memory are difficult to answer (Gray, 1955). There is now substantial evidence of 'forgetting': that memory declines with the length of the recall period, leading to under-estimation; see Sudman et al. (1996) for a review. The situation is complicated by the fact that forgetting does not occur at random but might be differential across respondents and types of questions.

A key development in the literature on recall expenditure questions was the identification of 'telescoping' as a significant problem by Neter and Waksberg (1964). This is the phenomena of respondents erroneously including in their response expenditures that occurred before the specified recall period, leading to an over-estimation of expenditure in the recall period. Neter and Waksberg documented this phenomena in the CE (particularly, home alterations and repairs). Telescoping is thought to arise because remembering dates is particularly difficult. This leads to an over-estimation of expenditure in the recall period since uncertainty over dates increases as one goes back farther in time. Thus, it is more likely for an older expenditure to be mistakenly placed in the recall period than it is for a more recent expenditure to be mistakenly placed prior to the recall period. This process has been formally modeled; see e.g., Rubin and Baddeley (1989). Recall answers could therefore be overestimated (because of telescoping) or underestimated (because of forgetting).

The earliest version (1960/61) of the CE had annual recall but this was abandoned for the 1972/3 survey because of the work of Neter and Waksberg (1964) and Sudman and Ferber (1971) on recall problems, particular telescoping.

Neter and Waksberg proposed ‘bounded’ recall as a way of minimizing telescoping problems. The idea is that the recall period should be marked by an interview to prevent prior expenditures entering the recall period. This suggestion has been adopted by the current design of the CE. The recall sample is interviewed five times with data from the first interview discarded; the first interview serves to mark the beginning of the first recall period. Data from the current CE is consistent with telescoping. For some categories of expenditure the (normally discarded) data from the first interview suggests significantly higher rates of expenditure for some categories of goods (Silberstein, 1990).

Problems with diary methods

In principal, a diary with perfect compliance and covering a sufficiently long period should give very good expenditure data. In practice, however, diary collection of expenditure information suffers from a number of problems.

First, respondents are typically asked to keep diaries only for short periods, partly in recognition that careful completion of a diary implies significant respondent burden. For categories of expenditure that are purchased irregularly, or at regular intervals that exceed the duration of diary keeping, infrequency problems will arise. This is a kind of measurement error: a household may over (or under) estimate their true rate of expenditure if the diary keeping period happens to include (or not include) a major shopping trip or a major purchase. While this may not affect estimates of average expenditure across households it certainly increases variance and will therefore bias estimates of inequality and poverty; it also causes bias when total expenditure is used as a ‘right hand side’ variable, as in the estimation of Engel curves.

A second concern is that compliance with diaries is certainly not perfect. In some budget surveys a great deal of diary completion occurs at the time of diary pick-up: interviewers collecting the diary check for completeness and often end up recording additional expenses. Silberstein and Scott

(1991) report that this occurs in as many as a quarter of CE diaries. In such cases the distinction between a diary survey and recall survey is not clear.

In addition, evidence from a number of diary surveys with two weekly diaries suggests that compliance declines with the duration of record keeping. Apparent rates of expenditure in the second week of diary keeping are lower, sometimes substantially so. In addition to the between week differences, within week responses tend to be significantly larger for the earlier days of either week. These patterns have been reported in the CE (Silberstein and Scott, 1991; Stephens, 2003), the Canadian Food Expenditure Survey (Statistics Canada, 1999; Ahmed et al., 2006) and the U.K. Family Expenditure Survey (Tanner, 1998). In the 1987 CE, expenditures in the second week of the diary were eleven percent below those in the first week (Silberstein and Scott, 1991). These patterns are typically attributed to 'diary fatigue' (for example, Silberstein and Scott, 1991, and Statistics Canada, 1999) and they have been known for a long time (e.g., Kemsley, 1961; Turner 1961; Sudman and Ferber, 1971; McWhinney and Champion, 1974).

An intriguing (and alarming) alternative explanation for the drop off in expenditure rates from first week diaries to second week diaries is that keeping a diary alters behavior. This would not be entirely surprising: in the popular personal finance literature making a record of expenditures is routinely advocated as a way of controlling expenditure and increasing saving. We identified only two studies, both from the U.K., that investigate this possibility. Kemsley et al. (1980) report on experiments with the U.K. Family Expenditure Survey. They conclude that behavioral responses to participation in the survey are not systematic or uniform. McKenzie (1983) is an early study of response problems with diaries, undertaken with the cooperation of British Telecom and based on telephone calls (where the diary record can be compared to metered usage). McKenzie concludes that there is no evidence in this study that keeping a diary affects telephone usage. Of course, this result does not necessarily generalize to other categories of expenditure.

Another kind of non-compliance with diaries is non-specificity (in which the respondent does not record sufficiently detailed information about a purchase). A closely related problem is that respondents sometimes record a single cost or expenditure for multiple items bought together. Silberstein and Scott (1991) report that seven percent of food purchases in the 1987 CE (totaling twenty-six percent of food expenditure) suffer from non-specificity. This non-specificity often implies that the data analyst needs to allocate non-specific expenditures to specific categories, which is a kind of imputation. Silberstein and Scott note these phenomena are much less of a problem in the interview survey, presumably because of the structure of the interview and the interaction with the interviewer.

A final concern with diaries is that they are expensive to administer. The way in which they are typically now used, with drop-off and collection and checking, involves multiple visits to the household.

Direct comparisons of diary and recall records

McWhinney and Champion (1974) report on early experiments in Canada that compared diary and recall methods of collecting expenditures. The conclusion of those studies was that *annual* recall (in conjunction with a cash-flow reconciliation or ‘balance edit’ – to be discussed below) gave data of good quality. The Canadian national budget survey (initially called the Family Expenditure Survey and later the Survey of Household Spending) maintained this design until very recently.

A number of recent studies have sought to compare diary versus recall methods, often for food expenditures. These studies exploit that the fact that a number of existing surveys, including the CE and the Canadian Food Expenditure Survey, ask respondents to estimate or recall usual food expenditures before subsequently completing a diary. This provides recall and diary measures for the same households. Using the Canadian data on food expenditure, Ahmed et al. (2006) show that recall and diary responses are different and the differences between them relate to both the level of expenditure and observable characteristics of the households. This implies that, perhaps unsurprisingly, there is nonclassical measurement error in one or both measures. Battistin and Padula (2010) show that recall and diary food expenditure measures are not rank preserving, meaning that recall and diary measures from the same household do not order those households by expenditure identically. This is important as rank preservation is among the weakest identification conditions required by econometric models of measurement error. Silberstein and Scott (1991) note that some categories of expenditure (e.g. apparel) exhibit different seasonal patterns in the diary and interview components of the CE.

Of course, it is insufficient to know that recall and diary measures differ; we would like to know which is superior. Recall measures almost always have a longer reference period, and hence will almost always have lower variance; the literature has not considered this a sensible criterion of comparison. Most categories of expenditure are thought to be under-reported, so that higher rates of expenditure have been taken to be indicative of less error.

In both the Canadian food survey (Ahmed et al., 2006) and the CE (Gieseman, 1987; Bee, Meyer and Sullivan, this volume) the recall measure of food is on average higher than the diary measure. This would be surprising if recall questions on food expenditure suffered from significant 'forgetting' and the diary records were accurate. Telescoping is unlikely to be the explanation for this finding as food expenditures are small and regular, and telescoping is thought to be a problem mostly for large and irregular expenditures. Diary fatigue and noncompliance may be an explanation. Statistics Canada apparently has greater confidence in the level of the recall measure as they routinely inflate the diary data to match the average of recall reports prior to release. Silberstein and Scott (1991) make comparisons for a number of items that are collected in both the interview and diary components of the CE. They report that the diary method produces higher expenditure estimates for some categories (apparel, home furnishings) while the interview produces higher expenditures for others (entertainment and hobbies.)

Gieseman (1987) compared food expenditure data from CE interview and Diary surveys and finds the former are significantly higher and closer to the PCE numbers from the National Income and Product Accounts (NIPA). Bee, Meyer and Sullivan, elsewhere in this volume, separately assess the CE interview and Diary surveys against PCE benchmarks derived from the NIPA for a range of expenditure categories. They report that for many large categories of expenditure the ratio of expenditure observed in the interview survey to the PCE benchmark is close to one, and they have not deteriorated over time; this is not true of the diary survey. See their chapter for additional details.

Mixed data collection methods

As noted in Silberstein and Scott (1991) many national budget surveys use a mix of diary and recall methods. The choice then becomes not which method to use for the survey but rather which method to use for each category of expenditure.[3] The considerations are similar to those just discussed. Detailed comparisons of expenditure reports in the two sources suggest that there may be some advantage in making these choices at very detailed item levels, but whether that advantage can be realized in practice is open to question. See Silberstein and Scott (1991) for further discussion.

Summary. There appears to be a common view that diary approaches provide more reliable measures of expenditure – it is almost a folk theorem that diary-based budget surveys set a 'gold standard' for measuring household expenditures. However, our review of the literature casts doubt on that conclusion. Response effects such as diary fatigue imply that diary-based measures are not necessarily error free, and since they clearly involve a much higher burden on respondents, selective participation might be a more severe concern than for the recall approach.

2.3 Response formats

With both recall and diary methods there are important questions of questionnaire design in general, and response format in particular. A key issue is whether one should employ open-ended (fill-in) formats or closed response formats such as range card or brackets. At least two aspects are important for this choice. The first is respondent burden – it is easier for the respondent to tick off

[3] Where both the diary and the interview components of the CE collect information on a category of expenditure, a similar decision is made in determining which data to use in the production of the integrated accounts.

one of a small number of specified ranges rather than provide a numerical estimate, so different response formats might result in different rates of item nonresponse. The second aspect concerns problems associated with each of the two formats: Open-ended questions typically yield rounded or heaped responses whereas closed formats might induce the respondent to use certain estimation strategies that produce systematic biases.

Pudney (2007) analyzed the responses to questions in the British Household Panel Study (BHPS) about spending on domestic energy (electricity, gas, etc.). He documented that responses are 'heaped' with large proportions of responses at particular 'focal' expenditure levels (i.e., prominent, round numbers such as multiples of 10, 50, or 100). Pudney argues that heaping results from the use of estimation strategies that involve rounding. Some respondents might choose a round number for weekly spending and then scale that up to an annual total, some use rounding at the monthly or annual level, while others do not round at all. (There might be an interesting interaction between rounding strategies and the choice of the reference period, another important aspect of questionnaire design which we review below.) These results suggest that rounding is differential across respondents. There is also some evidence from controlled experiments that the degree of response rounding is affected by the respondent's uncertainty about the target quantity (Ruud, Schunk, and Winter, 2011). Thus, simple strategies to correct for heaped responses in the analysis of the data that have been developed in the statistics literature (such as those that require a 'coarsened at random' assumption; Heitjan and Rubin, 1991) are too simplistic; see also Wright and Bray (2003).

One way to avoid the statistical problems associated with heaped responses is to use closed response formats that provide respondents with a list of brackets (a 'range card') from which they choose one. Another advantage of closed response formats is that they tend to produce lower rates of item nonresponse. But the data obtained from such bracketed questions also come with their problems – when the object of interest is a continuous and cardinal variable, information is lost and regression models require stronger assumptions compared to those that could be estimated with the continuous variable. Moreover, Manski and Tamer (2002) illustrate that these assumptions must be strong since the bounds on parameters of interest that can be identified from bracketed data are large.

Winter (2002), building on work in survey research and social psychology (Schwarz et al. 1985) shows that in addition to these statistical problems, bracketed data might introduce additional systematic biases. In a controlled survey experiment, he assigned respondents either to an open-ended or to three versions of bracketed questions that used different bracket thresholds; the target quantities were six expenditure items. The four question types delivered response distributions that are statistically different from each other. The response patterns are consistent with psychological theories of response behavior that predict that respondents who are uncertain about their response (here, their true expenditure on an item) use the information provided by the bracket thresholds to determine what the distribution of the target quantity in the population is and then give a relative response. For instance, a person who thinks her consumption is 'average' might tick of the middle category of a range card irrespective of the thresholds used. The biases that arise from such behavior can be large, and they are likely differential across survey respondents.

Similar systematic biases arise when follow-up bracketed questions (sometimes known as 'unfolding brackets') are used when respondents give item nonresponse to open-ended questions; e.g. van Soest and Hurd (2008). The underlying psychological mechanism in unfolding questions that require yes-no responses at each step (anchoring) is, however, slightly different from the one that affects range-card type questions (estimation and response on a relative scale).

There are a number of further issues in the design of diaries, including whether to pre-print expenditure categories on the diary, and whether diaries should be organized chronologically (as a 'journal') or by product or by outlet type. Indeed, Silberstein and Scott (1991) argue that questionnaire design issues are likely to be more important with diaries than with recall interviews, because of the absence of an interviewer who can help with survey comprehension and check for obvious reporting errors. Sudman and Ferber (1971) reported higher expenditure reports in diaries organized by product type in an experimental comparison with journal and outlet formats. Tucker (1998) and Tucker and Bennett (1988) report that preprinting expenditure categories in diaries leads to higher expenditure totals.

Summary. Both open-ended and closed response format recall questions produce data that are coarsened in non-random ways. Thus, they cannot be used with standard regression approaches and technically, they do not point-identify the parameters of interest (and if bounds are identified, they tend to be wide). It is an open question of whether the biases in open-ended or closed (bracketed) questions are larger. Leaving this choice aside, reducing the respondent's uncertainty about the quantity of interest by appropriate survey design should reduce the response biases and subsequent statistical problems associated with both response formats: Respondents who are less uncertain are less likely to use biased estimation strategies when they form their response, an issue to which we return below.

2.4 Disaggregation of expenditure categories

The issue of how finely survey instruments should disaggregate the components of quantities such as income or expenditure has been studied for a long time. In the following discussion, we focus on a situation in which the researcher is interested in getting an accurate measure of a quantity at an aggregate level, such as total expenditure on non-durable goods in a certain period. If a researcher has substantive interest in a variable at more disaggregate levels, such as food expenditure, that places a natural restriction on how much the components can be aggregated.[4]

Much of the early work on disaggregation we are aware of looks at income rather than consumption questions. Herriot (1977) compared four questionnaire variants and found that the more aggregated the income categories are, the less complete is the reporting of income. More recently,

[4] There has been work on "one-shot" questions about total expenditure of a household, particularly for use in general purpose surveys. As our focus here is on dedicated expenditure surveys, and any such survey will surely wish to capture more disaggregated detail, we do not review that literature here. See Browning et al. (2003) for an introduction.

Micklewright and Schnepf (2010) investigated the reliability of single-question measures of income. They compared the distributions of income in two U.K. surveys—individual income in the Office for National Statistics's Omnibus survey and household income in the British Social Attitudes survey—with those in two other surveys that measure income in much greater detail. They found that the distributions of single-question and more detailed measures compare less well for household income than for individual income.

There has been work on expenditure categories in both developed and developing countries. Joliffe (2001) reports findings from a survey experiment conducted in El Salvador. Longer, more detailed sets of questions resulted in an estimate of mean household consumption that was 31 percent greater than the estimate derived from a condensed version of the questionnaire, and the distributions of household consumption from the long and short questionnaires were also different. Joliffe further shows that the differences in estimated consumption lead to different substantive conclusions about levels of poverty in the population. Pradhan (2009) analyzes data from an experiment that occurred in a national household survey in Indonesia: Questions on consumption were asked with different levels of aggregation, and households were randomly assigned to the different designs. Like Joliffe, Pradhan finds that the level of aggregation has a significant effect on the estimate of total consumption.

Turning to expenditure surveys in developed countries, an early study by Reagan (1954) of farm operators found that total expenditure was only about 10% lower with 15 categories than with over 200. Winter (2004) conducted an experimental study with a large, representative sample in a Dutch internet panel survey (the CentERpanel). Respondents were randomly assigned to either a one-shot question on total monthly nondurables expenditure or to in a table with 35 disaggregated categories they had to fill in. The two designs produced significantly different distributions of the totals. Moreover, these differences varied with household characteristics. Underreporting was high for the middle income groups and decreased with income. Also, underreporting appeared to be most severe for middle-aged respondents. The findings are consistent with older households' nondurables expenditures being concentrated on few items and therefore easier to recall. Also, and perhaps not surprisingly, underreporting in the one-shot question is smaller for respondents who list 'housekeeper' as their occupation.

Focus group results reported by d'Ardenne and Blake (2011) suggest that respondents consider more disaggregated designs to not only a heavier burden but also more intrusive. This finding may be quite important in some settings or for particular sub-populations of households.

Summary. There are several studies that investigate the effects of disaggregation on survey measures of both income and expenditure. These studies suggest that designs that use more disaggregated categories yield higher estimates of the totals, presumably because households do not include some categories in their estimates of totals in one-shot or highly aggregated designs. It is not certain that greater disaggregation always leads to better results, particularly as respondents find more disaggregate demands more intrusive and a greater burden. It is worth noting that most of the studies cited above compare treatments all of which have less disaggregation than the current CE. There is also evidence (for example, Hurd and Rohwedder, this volume) that less disaggregated collection can capture many of the important life-cycle and time-series patterns of expenditure. It may well be that for *research* purposes, a less disaggregated design is sufficient. Finally, even if designs with more questions on more disaggregated categories yield better results, in practice there is still a trade-off between respondent burden and survey cost, and response quality. We are not aware of studies that try to quantify this trade-off and find optimal levels of disaggregation under a survey cost or time constraint.

2.5 Predicting aggregates from components or other variables

Given that measuring the aggregate quantity of interest (say, total household expenditure) using a one-shot question might provide unreliable results, and that asking for a longer list of components might not be feasible, an alternative approach is to ask questions on fewer expenditure items and employ them to predict the aggregate quantity using a statistical model. This model would be estimated using a separate, more detailed survey with reliable data on a large number of categories (typically, a household budget survey based on diaries); the estimated coefficients could then be used to predict the aggregate with a subset of the items in another survey. The statistical goal would be to have an unbiased prediction that preserves patterns of variance and covariance. The classic paper in this vein is Skinner (1987) on imputing total consumption of PSID respondents, on the basis of the limited expenditure questions in the PSID. There have been a number of proposed refinements to this procedure; recent examples are Blundell et al. (2008); Blundell and Pistafferi (2003); Battistin et al. (2003). Browning and Crossley (2009) propose a method by which moments of the total expenditure distribution can be recovered from information on just two goods; see Attanasio et al. (this volume) for an application of this method. Note, however, that these methods all require that information on the relationship between total expenditure and expenditure on categories of goods and services (that is, Engel curves) is available from some other source.

An alternative is to use the intertemporal budget constraint to impute consumption expenditure from data on income and wealth: Browning and Leth Peterson (2003) report one attempt to this with Danish data. Interestingly, recent U.K. focus group evidence (d'Ardenne and Blake, 2011) suggests that, when asked a question on total expenditure, many (but by no means all) respondents work out an answer by beginning with income and adjusting for changes in assets (primarily by subtracting savings.) The same focus group evidence suggests, though, that using survey questions on income and wealth changes to get total expenditure is unlikely to be a full solution, for a number of reasons. One problem identified in the focus groups is that respondents whose expenditures exceed their incomes find questions about changes in wealth very intrusive.

Summary. Our conclusion is that prediction of expenditure from components or from income and wealth data may be useful in particular contexts, but is not likely to be a major component to any replacement of current national budget surveys. The methods that use components to predict total expenditure require the existence of a budget survey for calibration, and methods based on income and wealth changes are very intrusive for significant subpopulations. Moreover, these methods do not capture the disaggregated spending information necessary for price index construction and many research applications.

2.6 Defining the response unit (and choosing the respondent(s))

Another fundamental design choice for expenditure surveys is: Should we measure household or personal expenditure? This question has various aspects. First, some expenditures arise only at the household level (such a rent, heating) and cannot be easily assigned to individual members; others are typically made at the household level but could, in principle, be assigned to individual members, such as many items purchased during regular trips to the grocery store; and yet others are made individually or can be assigned easily to individuals, such as clothing. Capturing these structures is difficult at the conceptual level and highly expensive to implement in interview surveys since different parts of the instrument would have to be assigned to the members of the household. Even if we aim to collect only the aggregate expenditure of the all household members on each item of interest, there remains the practical question of how to collect this information.

In many existing recall expenditures surveys, expenditure questions are given only to one respondent (typically, the person most knowledgeable about household finances) who is asked to provide estimates ‘for the household’. This can lead to two types of problems. First, great care must be taken in communicating the spending about which the question asks. The concept of a household – which economists often do not care to define in plain language, presumably because it is so natural to us – might be misunderstood. Respondents may report individual expenditure even when a question asks about household expenditure: Comerford, Delaney and Harmon (2009) provide experimental evidence on this problem. d’Ardenne and Blake (2011) report focus group evidence of a different misunderstanding: respondents believe that ‘household spending’ or even ‘spending of your household' means *only* shared expenses, or expenditures on those goods and services necessary to 'run the household.'

The second kind of problem is that that even the member of the household with the best knowledge of household finances may not know or be able to estimate the spending of other members. This can be interpreted as a proxy interview problem, and is likely to be particularly problematic in complex households: those with unrelated adults ('sharers') or multiple generations of adults. Browning, Crossley and Weber (2003) report evidence that non-response to household expenditure questions is much higher for such households. However, it is likely to pose difficulties for all types of households, apart from single person households. Focus group results reported in d'Ardenne and Blake (2011) confirm this conjecture and also suggest that this problem may be the more severe the finer the detail to be collected. Individual household members may be able to estimate the total spending of other household members but unable to provide much information on how that spending is broken down by goods and services.

The corresponding issue in diary surveys is how many diaries should be completed. The current CE design involves a single diary for the household, but some national budget surveys (U.K., France, Denmark) have multiple diaries (one for each household member above a minimum age). The choice of one or multiple diaries has been studied (Kemsley and Nicholson, 1961; Grooteart, 1986) and the evidence is mixed. Multiple diaries give higher totals, suggesting that some expenditures are missed with a single diary, but multiple diaries lead to a higher incidence of non-cooperation. Similar findings emerged from a small feasibility study commissioned by the CE in 2006 (Goldenberg and Ryan, 2009).

Beyond these difficulties with collecting household level expenditures, it is undoubtedly the case that the intra-household allocation of goods and services to individuals is of considerable interest to researchers and policy makers.[5] Individual diaries (or individual recall interviews), do not necessarily identify individual consumption. We do not know, for example, if one adult's expenditures are for themselves, for another adult, for children in the household, or to be shared. Bonke and Browning (2009) report on successful Danish experiments with collecting individual consumptions in household surveys by asking 'for whom?' in addition to the standard information collected on each expenditure item.

Summary. Asking one respondent, even the 'person most knowledgeable', to report expenditures leads to a number of possible response problems and errors. More detailed collection of data on expenditures made by different household members is potentially expensive. But where it is feasible, it may lead to higher quality data. If it can also be combined with data on who benefited from the expenditure, it opens up rich possibilities for studying allocations within households.

2.7 Reference period

Another fundamental issue in the design of survey instruments that elicit flow variables such as consumption or income is the choice of the reference period. Should we ask respondents for daily, weekly, monthly, or annual amounts? Is the optimal reference period different when measuring income and expenditure? Are there perhaps also differences in optimal reference periods across different expenditure items? Then, whatever the choice of reference period may be, should we ask respondents to provide reports for the past period or for a typical period?

The discussion in Section 2.2 above suggests that designers of recall questions face a tradeoff. Longer periods may lead to greater 'forgetting' and hence under-reporting. Shorter recall may generate measurement error through the infrequency of purchases. Because diary fatigue seems to lead to decreasing compliance throughout the recording period, designers of diary surveys face a tradeoff not unlike that faced by designers of recall surveys. Shorter recording periods will lead to

[5] See Deaton (1997), Section 4.3 and the references therein, for an introduction to the literature on intrahousehold allocation.

less bias in the estimation of mean expenditures, but, because of infrequency, higher variance. Infrequency will also lead to bias in estimates of dispersion. Longer recording periods will reduce infrequency problems but lead to greater underestimation. There is no reason for diary fatigue and 'forgetting' to follow the same time path, so that even for a given good, the optimal reference periods might also differ between recall and diary approaches.

At least in a recall survey it is quite feasible to vary reference periods by category of expenditure, and it seems obvious that the optimal reference period will be different for different categories of expenditure. Rates of forgetting depend on the frequency and on the salience of purchase (Silberstein and Jacob, 1989) which will also differ by category of expenditure.

Bradburn (2010) provides an excellent review of the cognitive processes that occur when survey respondents are asked to recall quantities from memory and maps theses processes into the issue of optimal lengths of recall periods. A central conclusion he draws is that 'no single recall period will be optimal for all events', but also that there is no general knowledge on what recall periods should be used for which goods, and that 'more empirical work is needed to determine the optimum recall periods for different categories of expenditures' (p. 8). Bradburn also discusses how questions with different recall horizons should be grouped within a questionnaire.

Clark, Fiebig, and Gerdtham (2008) present an interesting approach to estimate the optimal length of recall periods from prior survey data; their application is, however, not expenditure but the frequency of doctor visits and medical treatments during defined past periods.

Hurd and Rohwedder (2009) report evidence from controlled experiments on the tension between asking about spending over long and short time frames. They conclude that respondents' choice of reference period is related to their household's frequency and level of spending in a particular category. Respondents tend to choose a longer reference period for less frequently purchased items. Also, recall bias is important when using longer reference periods such as 'last 12 months'. They argue that longer reference periods should be used sparingly with relatively frequently purchased items. Finally, they confirm that short reference periods might provide an unrepresentative snapshot of household spending because of infrequent purchases. In the Consumption and Activities Mail Survey, a component of the Health and Retirement Study (HRS) they adopted an innovative alternative approach that allows respondents to choose from a set of reference periods of different lengths for each item.

Despite the theoretical considerations and evidence just described, not all evidence points to the desirability of different reference periods for different categories of expenditure. The Indian National Sample Survey Organization (NSSO) conducted a detailed experiment with different recall periods using daily visits as a gold standard measure. The study found no uniformly optimal recall length across all goods but a thirty day recall period (which was the baseline design) seemed to do reasonably well (Deaton and Kozel, 2005). There is also some suggestive evidence (McWhinney and Champion, 1974; see also the discussion in Deaton and Grosh, 2000) that annual recall works well, at least in some contexts.

A related issue is whether – given a period length – recall questions should be asked for the last period or for a typical period; the tradeoff being between recall accuracy (better for the most recent period) versus missing infrequent expenditure (which will be avoided when asking for a typical period). There is a related literature on measuring the frequency of regular behaviors. Chang and Krosnik (2003), for example, study survey questions on the frequency of news media consumption. They find that ‘typical week’ questions perform better than ‘last week’ questions in that context, but they also conclude that more systematic research is needed on how questions on the frequency of behaviors should be asked in other contexts.

With respect to expenditures, Edgar (2009) reports a cognitive interviewing study (76 participants) which examines four questions about ‘usual’ spending in the CE interview survey: food at home, food away, alcohol at home, and alcohol away. The study revealed a great deal of heterogeneity in the estimation strategies that respondents employed to answer these questions. Respondents seemed to interpret the term ‘usual’ in a variety of ways.

Angrisani and Kapteyn (this volume) designed and fielded an experimental module in a U.S. internet survey (the American Life Panel) in which they asked individuals to report the frequency of their purchases and the amount spent by debit cards, cash, credit cards, and personal checks. The data show that the type – specific or typical – and length of recall periods greatly influence household reporting behavior.

Summary. Different reference periods lead to significant differences in the distribution of responses. As noted in Deaton and Kozel (2005) these difference can in turn lead to dramatic differences in objects of interest like poverty rates. Theoretical considerations and some evidence suggest different reference periods for different categories of expenditure, although some of the evidence we have suggests a uniform reference period may work fairly well. Given the potential importance of design choice, further evidence would be welcome.

There is good evidence that the choice of usual (or typical) versus most recent period has a significant effect on the responses received. But it is not clear, from the evidence we reviewed, which approach is preferable.

2.8 The role of incentives

Incentives affect survey response behavior. In a standard neoclassical view of the survey respondent, incentive payments compensate the respondent for the opportunity cost associated with answering the survey. There is, however, also a principal-agent problem: Since the survey agency cannot observe the true response, the respondent generally has an incentive to provide too little effort – i.e., not to think as hard about the responses as he might. In a series of papers, Philipson (1997, 2001), Philipson and Lawless (1997) and Philipson and Malani (1999) pursue this view both using theoretical models and data from controlled experiments. A general finding of these studies is that measurement error is elastic with respect to the incentive paid, which opens up the possibility of optimally assigning incentive payments to different (groups of) respondents should appropriate conditioning variables be available in sample frame data; however, these ideas have, to our knowledge, not been pursued.

In the context of expenditure surveys, Kemsley and Nicholson (1960) report on a small experiment in which modest cash incentives raised the cooperation rate among households asked to complete a one week expenditure survey by fifteen percentage points. Sudman and Ferber (1971) report on an experiment with small gifts (a flag or a notebook) in the context of a diary-based household expenditure survey. They report that households receiving a gift are significantly more likely to cooperate with the survey and report higher expenditures. Ferber and Sudman reviewed these and several other small studies in the mid 1970s and concluded that the effects of financial incentives in expenditure surveys had not been well-studied at the time of their review (Ferber and Sudman, 1974).

The CE itself conducted experiments with monetary incentives in both the diary and interview surveys in 2005/06 (Goldenberg and Ryan, 2009). In both cases a quarter of respondent households received a twenty dollar debit card and a quarter received a forty dollar debit card. In the interview survey the forty dollar incentive improved response rates and a range of measures of data quality. The effects of the twenty dollar incentive were in most cases not statistically significant. In the diary survey, the incentives were less successful. They seemed to improve data quality but had little effect on response rates.

The most recent study which our review of this literature uncovered is Bonke and Fallesin (2010). These authors also show that incentives can increase cooperation of respondents in consumption surveys – in their specific application, they offered larger lottery prices for respondents who were willing to answer a survey over the internet (which as they argue is the more reliable mode) rather than over the phone.

Summary. The use of incentives in expenditure surveys seems to be an area where additional systematic research would be welcome. Current evidence suggests that incentives can improve the quality of data collected in expenditure surveys, but there is insufficient evidence to draw any conclusions on the optimal form or size of the incentives. A particularly interesting question is whether the optimal size of incentives varies with respondent characteristics, and whether it is possible to condition incentives on such variables to the extent they are known from the sampling frame.

2.9 Approaches to reduce or correct response errors in real time

Computer-assisted surveys (personal and telephone interviews as well as internet surveys) offer additional strategies for improving the reliability of consumer expenditure measurement.

A first approach is preloading of information. If data on income or assets are already available, either from earlier interviews or from earlier questions within an interview, these variables can be used to provide the respondents with cues or to check whether a response is reasonable. For instance, if a preload information says that disposable monthly income was $2000, and the respondent says that he spent $4000 on nondurable consumption items last month, he could be asked whether that amount is indeed correct. Carroll et al. (this volume) argue that the possibility of preloading information is one of the significant advantages of a longitudinal component in a budget survey. While such approaches can reduce the number of severe response errors and outliers, designing them involves some judgment and to the extent that preload information is itself unreliable, might even exacerbate response errors (e.g., Manski and Molinari, 2008; Bollinger and David; 2005).

The official budget surveys in Canada have long been based on an intensive interview, annual recall, and a field editing procedure in which budget balance is checked. Households that are too far ‘out of balance’ are asked to review expenditures, incomes and changes in money balances. This cash-flow reconciliation procedure, in fact, significantly predated the move to computer-assisted interviewing. The early 1960/61 CE had a similar balance edit (Deaton and Grosh, 2000) but when the survey was subsequently redesigned to address the research indicating problems with recall, the balance edit was dropped as being incompatible with the new design (that is, with a survey without annual recall).

Brzozowski and Crossley (2011) report some evidence on the efficacy of the balance edit in the Canadian survey. They exploit the fact that the balance edit was dropped from the survey design in one year, and then reintroduced the following year. Through comparisons to adjacent years, they show that the main effect of the balance edit appears to be in improving income reports, especially at the bottom of the income distribution.

Hurd and Rohwedder (2010) describe the use of something like a balance edit in the American Life Panel (ALP), which is an internet panel. They asked households to complete a monthly survey on 25 higher frequency purchase categories and a quarterly survey on 11 lower frequency categories. At the end of the survey, respondents were presented with a 'reconciliation screen' and asked to review and correct the information they had provided. Hurd and Rohwedder report that about 3% of entries were corrected and that this led to reductions in item nonresponse and in the size and frequency of outliers.

Fricker, Koop and Nhien (2012, this volume) report on a new experiment exploring the feasibility of a cash-flow reconciliation (balance edit) in a revised CE survey. See their chapter for more details. A key finding is that the reconciliation seems to improve responses even when income and expenditures and income are reported over different intervals.

Summary. Evidence from a variety of sources suggests that cash-flow reconciliations and other opportunities for respondents to review their answers improve data in budget surveys. This merits further study and consideration in budget survey design or redesign.

3. Where do we go from here?

Surveying this literature, we see three priorities for further research on the collection and analysis of expenditure data. First, while we are accumulating much evidence on the consequences of different design choices in expenditure surveys, we need a theoretical framework to organize and interpret this evidence. Second, we need to begin to think more explicitly about cost-benefit tradeoffs. Third, on the analysis side, we need approaches to the data that incorporate what we know about the nature of response behavior and measurement error into structural econometric analysis. We now discuss these three points in turn.

3.1 A conceptual framework for understanding response behavior

As we have seen, researchers and survey designers have collected considerable evidence on the effects of different design choices in the collection of expenditure information. To move forward, we need to place this evidence in a theoretical framework that allows us to understand the evidence, to guide future experimentation, and to offer at least tentative answers to counterfactual questions about survey design. This is a challenging prescription, but a conceptual model of the response process can be useful as a starting point.

The response process, as a source of measurement error, can be broken down in several distinct stages (or tasks), as in Figure 1. This schematic, adapted from Tourangeau, Rips, and Rasinski (2000), presents the standard conceptualization of the survey response process in psychology.[6] We have added aspects of response behavior in expenditure surveys in italics. Many of the sources of measurement error and consequences of design features outlined above fit naturally into this framework, and it seems to us the natural place to begin to develop a more theoretical perspective on the design of expenditure surveys.[7]

As one example of the utility of such a perspective, consider the puzzling evidence that annual recall may give higher quality data than shorter recall periods. A possible explanation (Deaton and Grosh, 2000), which the conceptualization in Figure 1 highlights, is that the lengthening the recall period changes the response strategy of the respondent from one of retrieval or ‘counting’ (with the

[6] See also Sudman et al., (1996). The literature contains a number of such schemes which are similar in their broad conceptualization of the response process but differ in some details.

[7] Hudomiet (2011) attempts to map such a conceptual model into hypotheses that are testable in a structural model of survey responses, and he presents some preliminary estimates using data from the Health and Retirement Study (HRS).

attendant problems of telescoping and forgetting) to a strategy of estimation (based on partial retrieval and other salient information). The respondent's estimation strategy may work well – as well, for example, as diaries, or bounded recall designs. At the same time, the literature on the psychology of social response suggests that where respondents use an estimation strategy, the quality of responses may be quite sensitive to what information is available and salient. For this reason, it could be that the quality of annual recall data described in McWhinney and Champion (1974) may be quite sensitive to particular aspect of the survey design (such as the budget balance perspective imposed on both the interviewer and respondent by the balance edit in the Canadian surveys.)

Another example of the application of this kind of conceptualization of the response process as a series of cognitive tasks is the analysis of recall periods in Bradburn (2010). Bradburn uses a conceptualization the response process to highlight the key considerations in determining recall period length. In our view this is a fine example of how such questions should be approached.

3.2 Systematic discussions of survey costs

Collecting data on expenditures is expensive. For example, Deaton and Grosh (2000) note that the CE costs about five times as much per household as the Current Population Survey (the main income survey in the U.S.). In the literature survey in Section 2, there is useful evidence on almost all the aspect of expenditure design we might be interested. However, what is lacking, in almost all cases, is a systematic comparison of the benefits (in terms of increased reliability of the measures) and costs (monetary costs of administration, implicit costs arising from item or unit nonresponse and selection).

Groves (1989) provides a classic development of the survey cost versus survey error perspective. Manski and Molinari (2008) are among the few economists who take such a perspective to survey design. They argue that survey designers 'should use an explicit loss function to quantify the trade-off between cost and informativeness of the survey and aim to make a design choice that minimizes loss' (p. 264). The specific design problem they study is the use of 'skip sequencing' –

whether all respondents should be asked about an item of interest or only a subset which is determined conditional on earlier responses. The key problem is that skip sequencing reduces survey cost, but since conditioning variables might be mismeasured themselves, will also tend to increase survey error.

Given the high cost of expenditure surveys, it seems clear to us that more explicit discussion of the tradeoffs between cost and quality are needed.

3.3 Econometric models that reflect response behavior

Few studies try to take what we know about structure of measurement error and incorporate this knowledge in structural econometrics (see McFadden et al., 2006). Traditionally, measurement error was dealt with by making the assumption that it is classical, i.e. additive and uncorrelated with any other variable in the model of interest (Bound et al., 2001). This assumption is unrealistic for many variables that are measured in surveys, and in light of the evidence we reviewed in the previous section, it certainly does not hold for survey measures of consumption expenditure. Nevertheless, the assumption that measurement error is classical is often made since it makes the effects tractable, at least in textbook cases.[8] A more recent literature relaxes the assumption of classical measurement error, but its focus is on general results which do not depend on – or take advantage off – what we might now about the structure of measurement error. There are a few papers that are exceptions, and we think these papers lead us in a very useful direction.

Perhaps most relevant for the present research agenda is the paper by Battistin and Padula (2010) who suggest a way to obtain a superior measure of total expenditure at the household level. The methods developed in this paper exploit the structure of the CE (particular the multiple reports

[8] In the linear regression model, classical measurement error either leads to inflated variances of the estimated parameters if it affects the dependent variable or to a downward bias in the size of the estimated coefficients if it affects an explanatory variable. In nonlinear models, even for classical measurement error the predictions are not as clear-cut any more, and the effects are analytically intractable.

of expenditures available in the survey) in a sophisticated econometric framework. It is a model of how such work can be done.

Pudney (2007) and Ruud, Schunk, and Winter (2011), already mentioned above, model rounding strategies used by respondents when they answer open-ended survey questions; the implication of these papers is that a 'coarsening at random' assumption on rounded data should be replaced by a model that uses explicit knowledge of the process that generates rounding, for instance the fact that respondents who are more uncertain about an item might be more likely to round their response. Hoderlein and Winter (2010) study the effects of recall errors in a structural econometric model of household consumption. They show that non-classical measurement error related from recall errors in consumption, which is the dependent variable, can have grave consequences on model estimates, in contrast to the conventional wisdom which is based on the fact that classical measurement error in the dependent variable does not bias parameter estimates in a linear regression.

Papers such as these remind us that we need to do both things: get better data and make better use of the data we have (and better use of knowledge we have of the flaws in the data we have.) There should, in general, be more interaction between survey design and analysis methods (McFadden et al.; Browning and Crossley, 2009). This interaction of course must be mindful of the fact that these are general use surveys, and should not be tailored for any particular analysis. Nevertheless, we think the potential returns are large.

Those of us who both analyze household expenditure data, and think about how to collect it, are sometimes in the strange position. We worry that survey respondents may not be able to answer our survey questions, but the models we will use the data to estimate imply that they should be able to answer. The problem is symmetric. If we knew more about how households allocate resources over time and goods, we could design better questions. But equally, if we learn about how to ask better expenditure questions, this should also help us develop better models of consumer behavior.

The possibilities of two-way exchange between data development and model development seem to us very promising.

References

Ahmed, N., M. Brzozowski, and T. F. Crossley (2010): Measurement errors in recall food consumption data. Unpublished manuscript, McMaster University, York University, and University of Cambridge.

Attanasio, O., E. Battistin and H. Ichimura (2004): What Really Happened to Consumption Inequality in the U.S.? NBER Working Paper No. 10338.

Angresani, M. and A. Kapteyn (2011): Measuring household spending and payment habits: The role of 'typical' and 'specific' time frames in survey questions. Unpublished manuscript, RAND, Santa Monica.

Barrett, G., P. Levell and K. Milligan (2012): A Comparison of Micro and Macro Expenditure Measures Across Countries Using Differing Survey Methods. This volume.

Battistin, E., R. Miniaci, and G. Weber (2003): What do we learn from recall consumption data? *Journal of Human Resources*, 38(2), 354–385.

Battistin, E. and M. Padula (2009): Survey instruments and the reports of consumption expenditures: Evidence from the consumer expenditure surveys. Unpublished manuscript, University of Padova and University of Venice.

Bee, A., B. Meyer and J. Sullivan (2012): Micro and Macro Validation of the Consumer Expenditure Survey. This volume.

Blundell, R. and L. Pistaferri (2003): Income Volatility and Household Consumption: The Impact of Food Assistance Programs. *Journal of Human Resources*, 38, 1032-1050.

Blundell, R., L. Pistaferri, and I. Preston (2008): Consumption inequality and partial insurance. *American Economic Review*, 98, 1887–1921.

Bollinger, C. R. and M. H. David (2005): I didn't tell, and I won't tell: Dynamic response error in the SIPP. *Journal of Applied Econometrics*, 20, 563–569.

Bonke, J. and M. Browning (2009): The Allocation of Expenditures within the Household: A New Survey. Fiscal Studies, 30, 461-481

Bonke, J. and P. Fallesen (2010): The impact of incentives and interview methods on response quantity and quality in diary- and booklet-based surveys. *Survey Research Methods*, 4, 91–101.

Bound, J., C. Brown, and N. Mathiowetz (2001): Measurement error in survey data. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, 3705–3843. Amsterdam: Elsevier.

Bradburn, N. M. (2010): Recall period in consumer expenditure surveys program. Unpublished manuscript, NORC, University of Chicago. Available online at http://www.bls.gov/cex/methwrkshp_pap_bradburn.pdf (last accessed March 8, 2012).

Browning, M. and T. F. Crossley (2009): Are two cheap, noisy measures better than one expensive, accurate one? *American Economic Review, Papers & Proceedings*, 99(2), 99–103.

Browning, M., T. F. Crossley, and G. Weber (2003): Asking consumption questions in general purpose surveys. *Economic Journal*, 113, F540–F567.

Browning, M. and S. Leth-Petersen (2003): Imputing consumption from income and wealth information. *Economic Journal*, 113, F282–F301.

Crossley, T. F., and M. Brzozowski (2011): Measuring the Well-being of the Poor with Income or Consumption: A Canadian Perspective. *Canadian Journal of Economics*. 44(1):88-106.

Chang, L. and J. A. Krosnick (2003): Measuring the frequency of regular behaviors: Comparing the 'typical week' to the 'past week'. *Sociological Methodology*, 33, 55–80.

Clarke, P. M., D. G. Fiebig, and U.-G. Gerdtham (2008): Optimal recall length in survey design. *Journal of Health Economics*, 27, 1275–1284.

Comerford, D., L. Delaney, and C. Harmon (2009): Experimental tests of survey responses to expenditure questions. *Fiscal Studies*, 30(3/4), 419–433.

d'Ardenne, J. and M. Blake (2011): Developing expenditure questions: Findings from focus groups. Technical Report, National Centre for Social Research (NatCen), London

Deaton, A. (1997): *The Analysis of Household Surveys*. Baltimore, MD and London, U.K.: Johns Hopkins University Press.

Deaton, A. and M. Grosh (2000): Consumption. In M. Grosh and P. Glewwe (Eds.), *Designing Household Survey Questionnaires for Developing Countries: Lessons from Ten Years of LSMS Experience*, chapter 17. Washington, DC: The World Bank.

Deaton, A., and V. Kozel (2005): Data and Dogma: The Great Indian Poverty Debate. World Bank Research Observer, 20(2), 177-199

Edgar, J. (2009): What does 'usual' usually mean? Unpublished manuscript, Bureau of Labor Statistics.

Essig. L. and J. Winter (2009): Item nonresponse to financial questions in household surveys: An experimental study of interviewer and mode effects. *Fiscal Studies*, 30, 367–390.

Ferber, R., and S. Sudman (1974): Effects of Compensation in Consumer Expenditure Surveys. *Annals of Economic and Social Measurement*, 3(2), 21-34.

Fricker, S., B. Kopp, and N. To (2012): Exploring the feasibility of implementing a cash-flow reconciliation approach in the consumer expenditure interview survey. This volume.

Gieseman, R. (1987): The Consumer Expenditure Survey: Quality Control by Comparative Analysis. *Monthly Labor Review*. 8-14.

Goldenberg, K and J. Ryan (2009): Evolution and Change in the Consumer Expenditure Surveys: Adapting to Meet Changing Needs. Mimeo.

Gray, P. G. (1955): The memory factor in social surveys. *Journal of the American Statistical Association*, 50, 344–363.

Groves, R. M. (1989): *Survey Errors and Survey Costs*. New York, NY: Wiley.

Hausman, J. A., W. K. Newey, and J. L. Powell (1995): Nonlinear errors in variables estimation of some Engel curves. *Journal of Econometrics*, 65, 205–233.

Heitjan, D. F. and D. B. Rubin (1991): Ignorability and coarse data. *Annals of Statistics*, 19, 2244-2253.

Herriot, R.A. (1977): Collecting Income Data on Sample Surveys: Evidence from Split-Panel Studies. *Journal of Marketing Research*, 14(3), 322-329

Hoderlein, S. and J. Winter (2010): Structural measurement errors in nonseparable models. *Journal of Econometrics*, 157, 432–440.

Hudomiet, P. (2011): Cognition and survey behaviour: Evidence from a validation study of earnings. Unpublished manuscript, University of Michigan.

Hurd, M. and S. Rohwedder (2009): Methodological innovations in collecting spending data: The HRS Consumption and Activities Mail Survey. *Fiscal Studies*, 30(3/4), 435–459.

Hurd, M. and S. Rohwedder (2010): The Effects of the Financial Crisis and the Great Recession on American Household. NBER Working Paper No. 16407.

Income Statistics Division, Statistics Canada, (1999): *1996 Food Expenditure Survey, Public-use Microdata Files.* Statistics Canada: Ottawa.

Jolliffe, D. (2001): Measuring absolute and relative poverty: The sensitivity of estimated household consumption to survey design. *Journal of Economic and Social Measurement*, 27(1-2), 1–23.

Kemsley, W. F. F. and J. L. Nicholson (1960): Some Experiments in Methods of Conducting Consumer Expenditure Surveys. *Journal of the Royal Statistical Society, Series A*, 123(3), 307-328.

Kemsley, W. F. F. (1961): The Household Expenditure Enquiry of the Ministry of Labour: Variability in the 1953-1954 Enquiry. *Journal of the Royal Statistical Society, Series C,* 10(3) 117-135.

Kemsley, W. F. F (1965): Interviewer Variability in Expenditure Surveys. *Journal of the Royal Statistical Society, Series A*. 128(1), 118-139.

Leicester, A. (2012): Using scanner data to construct detailed weights for certain categories of spending. This volume.

Manski, C. F. and F. Molinari (2008): Skip sequenzing: A decision problem in questionnaire design. *Annals of Applied Statistics*, 2, 264–285.

Manski, C. F. and E. Tamer (2002): Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2), 519–546.

McFadden, D., A. Bemmaor, F. Caro, J. Dominitz, B. Jun, A. Lewbel, R. Matzkin, F. Molinari, N. Schwarz, R. Willis, and J. Winter (2005): Statistical analysis of choice experiments and surveys. *Marketing Letters*, 16(3-4), 183–196.

McKenzie, J. (1983): The accuracy of telephone call data collected by diary methods. *Journal of Marketing Research*, 20, 417–427.

McWhinney, I. and H. Champion (1974): The Canadian experience with recall and diary methods in consumer expenditure surveys. In: S. V. Berg (ed.): *Annals of Economic and Social Measurement*, Volume 3, Number 2, 113–140. Cambridge, MA: National Bureau of Economic Research.

Micklewright, J. and S. V. Schnepf (2010): How reliable are income data collected with a single question? *Journal of the Royal Statistical Society A*, 173, 409–429.

National Sample Survey Organisation, Department Of Statistics, Government Of India (2000): *Choice of Reference Period for Consumption Data*. Report Number 447.

Neter, J. and J. Waksberg (1964): A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59, 18–55.

Philipson, T. (1997): Data markets and the production of surveys. *Review of Economic Studies*, 64(1), 47–72.

Philipson, T. (2001): Data markets, missing data, and incentive pay. *Econometrica*, 69(4), 1099–1111. Philipson, T. and T. Lawless (1997): Multiple-output agency incentives in data production: Experimental evidence. *European Economic Review*, 41, 961–970.

Philipson, T. and A. Malani (1999): Measurement errors: A principal investigator-agent approach. *Journal of Econometrics*, 91, 273–298.

Pradhan, M. (2009): Welfare analysis with a proxy consumption measure: Evidence from a repeated experiment in Indonesia. *Fiscal Studies*, 30(3/4), 391–417.

Pudney, S. (2007): Heaping and leaping: Survey response behavior and the dynamics of self-reported consumption expenditure. Unpublished manuscript, University of Essex.

Reagan, B.B. (1954): *Condensed versus Detailed Schedule for Collection of Family Expenditure Data*. Agricultural Research Service, U.S. Department of Agriculture.

Rubin, D.C. and A.D. Baddeley (1989): Telescoping is not Time Compression: a Model of the Dating of Autobiographical Events. *Memory and Cognition*, 17.

Ruud, P. A., D. Schunk and J. Winter (2011): Uncertainty and rounding in survey responses: A laboratory experiment. Unpublished manuscript, University of Munich.

Safir, A. and K. L. Goldenberg (2008): Mode effects in a survey of consumer expenditures. Unpublished manuscript, Bureau of Labor Statistics (BLS), Washington, DC. Available online at http://www.bls.gov/cex/cesrvymethssafir1.pdf (last accessed March 8, 2012).

Shin, E., T. P. Johnson, and K. Rao (2011): Survey mode effects on data quality: Comparison of web and mail modes in a U.S. national panel survey. *Social Science Computer Review*, forthcoming.

Skinner, J. (1987): A superior measure of consumption from the Panel Study of Income Dynamics. *Economics Letters*, 23, 213–216.

Silberstein, A.R. (1990): First Wave Effects in the U.S. Consumer Expenditure Interview Survey. *Survey Methodology*, 16:293-304.

Silberstein, A.R., and S.Scott (1991): Expenditure Diary Surveys and their Asssociated Errors, in Biermer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman, Editors, *Measurement Errors in Surveys*, Wiley, Hoboken NJ: 1991.

Stephens, M. (2003): ‘3rd of the Month’: Do Social Security Recipients Smooth Consumption Between Checks? *American Economic Review* 93(1), 406-22.

Sudman, S. and R. Ferber (1971): Experiments in obtaining consumer expenditures by diary methods. *Journal of the American Statistical Association*, 66, 725–735.

Sudman, S. and R. Ferber (1974): A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. *Journal of Marketing Research*, 11, 128–135.

Sudman, S., N. M. Bradburn, and N. Schwarz (1996): *Thinking About Answers*. San Francisco: Jossey-Bass.

Tanner, S. (1998): How Much Do Consumers Spend? Comparing the FES and National Accounts. in Banks, J. and P.Johnson, Editors, *How Reliable is the Family Expenditure Survey?* INstitue for Fiscal Studies, London: 1998.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000): *The Psychology of Survey Response*. New York, NY and Cambridge, U.K.: Cambridge University Press.

Tucker, C. (1992): The Estimation of Instrument Effects on Data Quality in the Consumer Expenditure Survey. *Journal of Official Statistics*, 8

Tucker, C., and C. Bennett (1988): Procedural Effects in the Collection of Consumer Expenditure Information. *Proceedings of the Section on Survey Research Methods,American Statistical Association.*

Turner, R. (1961): Inter-Week Variations in Expenditure Recorded During a Two-Week Survey of Family Expenditure. *Journal of the Royal Statistical Society, Series C*, 10(3), 136-146.

van Soest, A. and M. Hurd (2008): A test for anchoring and yea-saying in experimental consumption data. *Journal of the American Statistical Association*, 103, 126–136.

Winter, J. (2002): Bracketing effects in categorized survey questions and the measurement of economic quantities. Discussion Paper No. 02-35, Sonderforschungsbereich 504, University of Mannheim.

Winter, J. (2004): Response bias in survey-based measures of household consumption. *Economics Bulletin*, 3(9), 1–12.

Wright, D. E. and I. Bray (2003): A mixture model for rounded data. *The Statistician*, 52:1, 3–13.

Figure 1: Schematic of the survey response process

1. Comprehension

    → Identify question focus (information sought)

    → Link key terms to relevant concepts

    Description of the items

2. Retrieval or recall

    → Generate retrieval strategies and cues

    → Retrieve specific, generic memories

    → Fill in missing details

    Effects of the length of the recall period

    Number of categories asked (what we often call aggregation)

3. Judgement

    → Assess completeness and relevance of memories

    → Integrate material retrieved

    → Form estimate based on partial retrieval and other salient information

    Effects of brackets on response (range-card type and unfolding)

4. Response

    → Map judgment onto response scale

    → Edit response

    Nonresponse for sensitive items

Source: Tourangeau, Rasinsky, and Rips (2000, p. 8)