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: 1 December 2011
Very preliminary and incomplete, comments very welcome,
and please check for a new version before citing.

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 synthesize these findings and discuss how they may be aligned conceptual model of survey response behavior. 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. 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 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 US 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, this volume). There is also substantial evidence survey design and quality affects substantive conclusions about important research questions. A good example is the study of the evolution of inequality in the US by Attanasio, Battistin and Ichimura (2004).

These facts have led to a number of initiatives to 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.

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 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 necessarily selective and circumscribed. It is aimed primarily at economists and researchers that traditionally analyse expenditure data, but whom 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, discussing evidence from less developed countries when it seems to us particularly useful. Deaton and Grosh (2000) includes many more results from developing countries. Finally our discussion will inevitably reflect some of our own interests and prejudices. Nevertheless we hope this review provides a useful introduction to what is known about the survey design in this area.

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 when detailed recall questions are used; the definition of the response unit; the reference period; the role of incentives; and approaches to reduce or correct response errors in real time. We draw tentative conclusions on each of these issues.

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" parts personal interviews, the use of paper questionnaires has become rare.) 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.

Key aspects of the interaction between survey mode and response behavior that are relevant for consumption 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 pres-

ence on an interviewer might increase such concerns). Models of survey response behavior, reviewed below, suggest that written surveys enhance respondents' understanding of survey questions relative to oral presentations, while confidentiality concerns will be more relevant in personal than in self-administered interviews. While there is a large literature on such effects in survey research, there is little systematic evidence when it comes to asking for consumption expenditure. Moreover, 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.

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. In comparison to the computer-assisted personal interview (CAPI) mode, rates of non-response were lower in the paper-and-pencil (P&P) drop-off questionnaire that could be answered in private and independently of the rest of the survey interview. This effect was pronounced for all six asset categories they analysed, while there was no item non-response significant effect for 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 covered both consumption and time use questions. Overall, they found that

response quality was higher when respondents chose to participate in the internet survey relative to participating in the telephone interview. Due to the self-selection of respondents, the mode effect cannot however be interpreted causally.

A consistent finding of many studies is that response rates to simple expenditure questions are higher than response rates to comparable income questions (Browning, Crossley, and Weber, 2003) suggesting that respondents view expenditure questions as being less sensitive; this has been corroborated in the recent UK focus group studies summarized by d'Ardenne and Blake (2011).

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 consumption 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 expenditure is the use of scanner data; this approach is covered in detail by Leicester (2012, this volume) and therefore not discussed here. For each of these strategies, 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 sec-

tions; in this section we review evidence that concerns the choice between recall and diary approaches.

Problems with Recall

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 subtantial evidence of "forgetting": that memory declines with the length of the recall period, leading to under-estimation. Sudman et al. 1996 for a review.

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 the phenomena of respondents erroneously including expenditures that before the specified recall period in their response, leading to an over-estimation of expenditure in the recall period. Neter and Waksberg documented this phenomena in the CEX (particularly, home alterations and repairs). Telescoping is thought to arise because remembering dates is particularly difficult. The reason that trouble remembering dates leads to an over estimation of expenditure in the recall period is that uncertainty over dates increases as one goes back farther in time. This means 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 eg., Rubin and Baddeley (1989). Recall answers could therefore be overestimated (because of telescoping) or underestimated (because of forgetting).

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 CEX. 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 CEX is consistent with telescoping. For some categories of expenditure the (nor-

mally discarded) data from the first interview suggests significantly higher rates of expenditure for some categories of goods).

The earliest version (1960/61) of the CEX 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.

Problems with diaries

In principal, a diary with perfect compliance and covering a sufficiently long period should give very good expenditure data. In practice, however, there are 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 effect 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 (including the CEX: Deaton and Grosh, 2000) a great deal of diary completion occurs at the time of completion (the interviewer collecting the diary checks for completeness and often ends up recording additional expenses.) 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. This is true of the CEX [Reference], the Canadian Food Expenditure Survey

[Ahmed et al., 2006] and the UK Family Expenditure Survey [REFERENCE]. It is also apparent in the trip diaries of the UK National Travel Survey. This pattern is typically attributed to "diary fatigue".

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 way of controlling expenditure and increasing saving. McKenzie (1983, J.Marketing Research) is the one study we are aware of that investigates this possibility directly. This is 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.

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 balance edit – to be discussed in Section 2.9 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 (which was also the design of the 1960-61 CEX) until very recently.

A number of recent studies have sought to compare diary versus recall methods particular food expenditures. These studies exploit that the fact that a number of existing surveys,

including the CEX 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, 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 measures not rank preserving.

Another point is that in both the Canadian food survey (Ahmed et al., 2006) and the CEX (Gieseman, 1987 check) 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 non-compliance may be an explanation (though in the Canadian data even the first week diary is on lower than the recall measure). Statistics Canada apparently has greater confidence in the level of the recall measure as, they routinely inflate the diary data to match average of recall reports prior to release.

Summary. There appears to be a consensus 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 consumption. However, some studies we reviewed cast doubts 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

This issue primarily applies to recall questions. When quantities such as income or consumption are measured using recall questions, an important design choice concerns the response format: Should we use open-ended (fill-in) 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 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 numbers of responses at particular expenditure levels. 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 the `coarsened at random' approach by 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 he chooses 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 questions 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).

One response format issue in the design of diaries is whether to pre-print expenditure categories on the diary. Tucker (1998; Journal of Official Statistics) 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 cannot be used with standard regression approaches since, technically, they do not identify the parameters of interest because the data are coarsened in nontrivial ways. 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 estimation problems with both response formats.

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.

Evidence on "one-shot" questions: see Browning, Crossley, Weber, 2003.

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, Micklewright and Schnepf (2010) investigated the reliability of single-question measures of income. They compared the distributions of income in two UK 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.

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 a quasi-experiment that occurred in a national household survey in Indonesia: Questions on consumption were asked with different levels of aggregation, and households are 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, a study by Regan (1954) of farm operators in which 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 charac-

teristics. 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.

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. Even if designs with more questions on more disaggregated categories yield better results, in practice there is still a trade-off between respondents 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

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 linger 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 imput-

ing total consumption to the PSID, on the basis of the limited expenditure questions in the PSID. There have been a number of preposed refinements to this procedure; recent examples are Blundell, Pistafferi, Preseton AER; Blundell and Pistafferi JHR; Battitin Miniaci and Weber, JHR.

An alternative is to use the intertemporal budget constraint to impute consumption expenditure from data on income and wealth: Browning and Leth Peterson (EJ, 2003) report one attempt to this with Danish data. Interestingly, recent UK focus group evidence (d'Ardenne and Blake, 2011) suggests that many respondents when asked a question on total expenditure 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 changes wealth 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 intrusive.

Summary. To be completed.

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 item purchase 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.

In most existing 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 times 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 estimate the spending of other members. This is particularly problem in complex households: households 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 be pose difficulties for all types of households, apart from single person households. Focus group results reported in d'Ardenne and Blake (2011) confirm this and also suggest that this problem may be more severe the more 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.)

In addition to the problems associated with asking a single respondent to report household expenditure is the fact that household expenditure misses detail in intrahousehold allocations that may be of considerable interest. [Some references.]

While the CEX has one household survey, some national budget surveyssurveys (UK, Denmark) have multiple diaries, though this alone does 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, (Fiscal Studies, 2009) report on successful Danish experiments with collecting individual consumptions in household surveys by asking *for whom?*

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 compbined 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 should we have chosen not to use a one-shot question but more disaggregated designs? 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? It is obvious to us that along these dimensions, there are no designs that work equally well for different target quantities.

The discussion in Section 2.2 suggests that designers of recall questions face a trade-off. 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 less bias, but, because of infrequency, higher variance. Longer recording periods will reduce infrequency problems but lead to greater underestimation. Moreover, there is no reason for diary fatigue and "forgetting" follow the same time path, so that even for the a given good, the optimal reference periods might also differ between recall and diary approaches. There is little systematic knowledge about how these design choices should be made.

As noted above, there is 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.

Hurd and Rohwedder (2009) report evidence from controlled experiments on the tension between asking about spending using a 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 provided 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.

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 ex-

penditure but the frequency of doctor visits and medical treatments during defined past periods.

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). We are not aware of systematic evidence on this design choice for expenditure questionnaires, but there is a related literature in survey research; see Chang and Krosnik (2003).

- Focus group results (d'Ardenne and Blake)
- (Also Angrisani and Kapteyn, this volume)

Summary. *To be completed.*

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 paper, Philipson (1997, 2001), Philipson and Lawless (1997) and Philipson and Malani (1999) pursue this view both using theoretical models and data from controlled experiments.

In the context of an expenditure survey, Bonke and Fallesin (2010) 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. *To be completed.*

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 consumptions measures.
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 provide the respondents with cues or to check whether a response is reasonable. For instance, if a prelod information says that disposable monthly income was $2000, and the respondent says that we spent $4000 on nondurable consumption items last month, he could be asked whether that amount is indeed correct. While such approaches can reduce the number of severe response errors and outliers, designing them involves some judgment and to the extend that preload information is itself unreliable, might even exacerbate response errors (e.g., Manski and Molinari, 2009; 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 procedure, in fact, significantly predated the move to computer-assisted interviewing. The early 1960-61 CEX 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 (CJE 2011) report some evidence on the efficacy of the balance edit in the Canadian survey. They exploit the fact that in one year the balance edit dropped. 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 real time revisions on the Internet Panel

  Summary. To be completed.

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, 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. We have added aspects of response behavior in expenditure surveys in italics. Many of the sources of 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 orientation to the design of expenditure surveys.

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 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.)

3.2 Systematic discussions of survey costs

Collecting data on expenditures is expensive. For example, Deaton and Grosh (2000) note that the CEX 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, systematic comparisons of the benefits (in terms of increased reliability of the measures) and costs (monetary costs of administarion, implicit costs arising from item or unit nonresponse and selection.)

- Groves (1989) on the survey cost vs. survey error perspective,
- Manski and Molinari (2008) who are among the few economists who take such a perspective to survey design.

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., 2011). 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 household expenditure. Nevertheless, the assumption that measurement error is classical is often made since it makes the effects tractable, at least in textbook cases.[1] A more recent literature relaxes the assumption of classical measurement error, but 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 current conference is Battistin and Padula (2010), which suggest away to obtain a superior measure of total expenditure at the household levels. The methods developed in this paper exploit the structure of the CEX (particular the multiple reports 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. Hoderlein and Winter (2010) study the effects of recall errors in

[1] 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.

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 expenditure (or "consumption") 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 questions which the models we will estimate with the data imply they must 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, that 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.

- Tucker, Biemer and Meekins.

4. Conclusion

To come.

References (incomplete)

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 US? NBER working paper no 10338.

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.

Beegle, K., J. DeWeerdt, J. Friedman, and J. Gibson (2010): Methods of household consumption measurement through surveys: Experimental results from Tanzania. Policy Research Working Paper 5501, World Bank, Washington, DC.

Blair, E. and S. Burton (1987): Cognitive processes used by survey respondents to answer behavioural frequency questions. *Journal of Consumer Research*, 14, 280–288.

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 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.

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.

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, MDand London, UK: 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.

Gibson, J., J. Huang, and S. Rozelle (2003): Improving estimates of inequality and poverty from urban China's household income and expenditure survey. *Review of Income and Wealth*, 49, 53–68.

Gibson, J. and B. Kim (2007): Measurement error in recall surveys and the relationship between household size and food demand. *American Journal of Agricultural Economics*, 89, 473–489.

Gibson, J. and B. Kim (2010): Non-classical measurement error in long-term retrospective recall surveys. 72, 687–695.

Goode, A. (2007): Recall bias in reported events: HILDA,Waves 1–5. HILDA Project Discussion Paper Series No. 2/07, University of Melbourne.

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.

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

Hudomiet, P. (2011):

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.

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.

Lanjouw, J. O. and P. Lanjouw (2001): How to compare apples and oranges: Poverty measurement based on different definitions of consumption. *Review of Income and Wealth*, 47(1), 25–42.

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

Manski and Molinari (2008)

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.

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.

Moore, J. C., L. L. Stinson, and E. J. Welniak, Jr. (1999): Income reporting in surveys: Cognitive issues and measurement error. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, and R. Tourangeau (Eds.), *Cognition and Survey Research*, 155–174. New York: Wiley.

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.

Pischke, J.-S. (1995): Measurement error and earnings dynamics: Some estimates from the PSID validation study. *Journal of Business and Economic Statistics*, 13(3), 305–314.

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.

Rohwedder, S., S. J. Haider, and M. D. Hurd (2006): Increases in wealth among the elderly in the early 1990s: How much is due to survey design? *Review of Income and Wealth*, 52(4), 509–524.

Schwarz, N., H. J. Hippler, B. Deutsch, and F. Strack (1985): Response categories: Effects on behavioural reports and comparative judgements. *Public Opinion Quarterly*, 49, 388–395.

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

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.

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

Tucker, C., P. P. Biemer, and B. Meekins (2011): Estimating underreporting of consumer expenditures using Markov latent class analysis. *Survey Research Methods*, 5, 39–51.

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.

Wilkins, R. and C. Sun (2010): Assessing the quality of the expenditure data collected in the selfcompletion questionnaire. HILDA Project Discussion Paper Series No. 1/10, University of Melbourne.

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.

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)