Policy Research Working Paper 9255

The Insights and Illusions of Consumption Measurements

Erich Battistin
Michele De Nadai
Nandini Krishnan

WORLD BANK GROUP
Poverty and Equity Global Practice
May 2020

Policy Research Working Paper 9255

Abstract

Although household well-being is anchored in long-term average rates of consumption, welfare comparisons typically rely on shorter-duration survey measurements. This paper develops a new strategy to identify the distribution of these long-term rates by leveraging a large-scale randomization that elicited repeated short-duration measurements from diaries and recall questions. Identification stems from diary-recall differences in reports from the same household, does not require these reports to be error-free, and hinges on a research design with broad replicability. This strategy delivers cost-effective suggestions for designing survey modules to yield the most accurate measurements of consumption well-being, and offers new insights for interpreting and reconciling diary-recall differences in household expenditure surveys.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at nkrishnan@worldbank.org.

The Insights and Illusions of Consumption Measurements*

Erich Battistin†
University of Maryland, CEPR, FBK-IRVAPP and IZA

Michele De Nadai ‡
University of New South Wales

Nandini Krishnan§
World Bank

JEL Classification: C81, D31, D63, E21, I32

Keywords: Household Consumption, Household Expenditure Surveys, Inequality, Poverty and Household Well-Being, Modes of Data Collection

*This version Thursday 21st May, 2020. The manuscript benefited from comments by Kathleen Beegle, Leah Bevis, Joachim De Weerdt, Beatriz Godoy, Arthur Lewbel, Juan Munoz, Sergio Olivieri, Thomas Pave Sohnesen and Tara Vishwanath as well as seminar participants at the CEPR conference on New Consumption Data (September 2019), NEUDC (October 2019) and the IARIW-World Bank conference on "New Approaches to Defining and Measuring Poverty in a Growing World"(November 2019). The views expressed here are those of the authors alone.

†Department of Agricultural and Resource Economics, 7998 Regents Drive, Symons Hall, College Park, MD 20742, United States of America. Contact: ebattist@umd.edu.

‡UNSW Business School UNSW Sydney, NSW 2052, Australia. Contact: m.denadai@unsw.edu.au.

§Poverty and Equity Global Practice, World Bank, 1818 H Street, NW Washington, DC 20433, United States of America. Contact: nkrishnan@worldbank.org.

1 Introduction

Household surveys are the dominant tool to measure and monitor welfare in low- and middle-income countries. In these contexts, measures of material well-being typically consider a combination of household consumption and expenditures, of which food comprises a large share. Indeed, global monitoring of extreme poverty relies on food-based estimates of welfare, such as the Cost of Basic Needs approach. Proxy means-tested targeting of interventions also derives from these estimates of welfare. Many programs worldwide target food consumption and nutrition, ranging from school meals to food subsidies like the Supplemental Nutrition Assistance Program (SNAP). Because of its centrality to policy and welfare measurement, a large body of the literature has questioned the accuracy of different modes of collecting consumption data to assess household well-being. We revisit this issue at a time when developing countries have increasingly moved to recall modes to collect food consumption in their national expenditure surveys.

Acquisition diaries have become the workhorse for benchmarking comparisons of household measurements from alternative collection modes. Beegle et al. (2012) randomize households in Tanzania to individual assisted diaries to define a benchmark against alternative survey designs that use recall questions. This is possibly the most important experiment on consumption measurement in low- and middle-income countries in the last decade (a more recent study for Niger is Backiny-Yetna et al., 2017). The accuracy of consumption measurements is also of concern in the developed world. For example, Brzozowski et al. (2017) compare recall questions on food spending from the Canadian Food Expenditure Survey to data from expenditure diaries. Battistin and Padula (2016), among others, do so using the Consumer Expenditure Surveys in the United States. In these instances, too, the accuracy of reports is assessed against diaries.

The assumption that diaries provide an error-free benchmark has little statistical justification. Diaries arguably minimize various types of reporting errors, including recall errors, telescoping, rule of thumb, and errors of omission (Lyberg and Kasprzyk, 2004). However, diaries are far from perfect. Most diary surveys are characterized by differences in reporting within the interview period, usually with a declining pattern over time (Silberstein and Scott, 2004). Possible explanations are not only declining cooperation due to fatigue but also the tendency of participants to deviate from the usual purchasing behavior as an effect of the diary or social pressure (Peterson Zwane et al., 2011). The proxy reporting of events experienced by persons other than the respondent yields figures different from those obtained from individual diaries (Beegle et al., 2012). The ratio between food totals computed from household surveys using diaries and national accounts have been found to be considerably low (Gieseman, 1987, and Bee et al., 2013). Moreover, diaries are expensive, as completion involves a reasonable standard of literacy and commitment, and regular visits from interviewers to ensure quality standards boost costs.

Faced with high costs of implementation, respondent fatigue, the need for enumerator effort, and the risks of enumerator shirking, many developing countries have moved to collecting recall data.[1] In recall interviews, households are typically asked to report on food *consumption* during the specified period and

[1]The World Bank's Living Standards Measurement Study (LSMS) survey finder includes a list of household consumption and expenditure surveys: of almost 90 surveys from more than 25 countries, more than 75 use recall. See also the guidelines on the use of recall interviews endorsed by the United Nations Statistical Commission (FAO and The World Bank, 2018).

to value that consumption, irrespective of source, at market prices. Diaries instead collect information on food *acquired* during the interview period regardless of when the food was consumed. There is little evidence on the best approach for managing the transition to recall interviews, with most countries accepting a break in the series on poverty, food security, and welfare aggregates, given that the survey mode affects indicators of poverty and inequality in the cross-section (Beegle et al., 2012) and over time (Battistin, 2003, Pudney, 2008, Attanasio and Pistaferri, 2016, and Coibion et al., 2020).

This paper studies how to obtain the most accurate measurements of population well-being in household expenditure surveys. Specifically, we define a household's well-being as their *usual* (or *average*) rate of consumption in a fixed period, which in our application is one week. Usual consumption is a longer-run average of weekly intakes and the quantity grounding welfare comparisons on the theoretical concept of consumption smoothing. The empirical challenge then is to use survey measurements collected as a snapshot, typically covering one week, to infer usual consumption. We develop a strategy to identify the distribution of usual consumption, which is *unobserved* because of the short-term nature of interviews, without applying any corrections to raw data or assuming *ex ante* which mode of collection – diaries or recall questions – yields the most accurate measurement.

We frame the identification of consumer welfare in a general setting, encompassing "survey errors" – defined here as differences between survey measurements and usual consumption – related to household preferences and misreporting behavior. Specifically, households with different consumption preferences in the interview week but planning for the same dollar amount of consumption over a longer period (e.g., a month) will have the same well-being. Such *irregularity of consumption* has implications for welfare analysis even when survey measurements are not misreported (as discussed by Kay et al., 1984, and Deaton and Grosh, 2000, among others). In addition, our analysis allows for survey errors from any of two broad sources. The first source is *infrequency of acquisitions*, which arises because cash outlays occurred during the interview week may present extreme values resulting from no acquisition or positive acquisition of infrequently purchased items.[2] The second source of survey errors arises because of *misreporting*, which is the object of a voluminous literature spanning both developed and developing countries (see Deaton and Zaidi, 2002, and Carroll et al., 2015). Our investigation disentangles the relative contribution of any of the above channels in both diary and recall data.

We leverage a unique large-scale experiment on consumption measurement in Iraq designed for the Iraq Household and Socio-Economic Survey (IHSES) in 2012. The survey ran continuously for one year and was administered to approximately 25,000 households across the country. All households filled out a 7-day diary on their acquisitions with the assistance of enumerators during regular visits. One-third of households were randomized to an additional survey module, administered before the diary, that asked them to recall food consumption in the last 7 days. The size of the experimental sample, approximately $8,000$ households, is of a different order of magnitude compared to other studies on consumption measurement in developing countries. The household diary employed was of the standard acquisition type, with one respondent recording on behalf of the household. Diaries were filled out by the same person

[2]If stocks are observed, one can impute consumption from diaries by measuring stock inflows and outflows (as in Beegle et al., 2012). However, the measurement of stocks is costly, prone to additional measurement concerns (Sharp et al., 2019), and seldom available in empirical work.

answering the recall module. The latter module was specifically designed to inform the national statistics agency on the transition from diary to recall, which is planned for 2021.

Our methodological innovation is in the use of diary and recall measurements for the *same* household. Compared to prior work, we derive conditions – building on Hu and Schennach (2008) – to identify nonparametrically any functional of the underlying distribution of usual consumption allowing for errors in both survey measurements. Our strategy also yields identification of survey error distributions for recall and diary measurements without assuming that errors are mutually independent or independent of usual consumption.[3] Moreover, our estimation framework leaves unrestricted the functional forms of usual consumption and survey error distributions. The availability of repeated diary-recall measurements of food spending or acquisition is not unique to our setting, the Consumer Expenditure Surveys in the United States and the Canadian Food Expenditure Survey being notable examples. This substantiates the external validity of our approach in a number of contexts beyond the specific case study.

Among the conditions needed for identification, three play a fundamental role (as shown in Figure 3). First, we assume that household acquisition in the diary week is a random variable centered on household usual consumption. We make the theoretical case for this assumption using a model of consumer demand (as in Meghir and Robin, 1992). This model also sets the reasons for expecting good estimates of consumption *averages* from diary data, but not of inequality or poverty rates in general.[4] Second, we seek variability in survey measurements induced by a change in usual consumption. Specifically, we build on the IHSES sampling design and assume that household usual consumption covaries with the average acquisition of same-income households located in other census enumeration areas and interviewed the same week. Our key restriction is that the average acquisition of these peers correlates with household diary and recall measurements only through the household's own consumption. Third, while we allow survey errors in recalled consumption and diary acquisitions to be dependent, we channel their dependence through household usual consumption. This assumption is standard in nonparametric identification with measurement error (Schennach, 2013), and its interpretation along with possible threats to its validity are discussed using the consumer demand model above.

Our first key insight is that diary and recall reports are not rank preserving, meaning that they do *not* order the same household identically in the data. We rule out simple explanations for this result. For example, we show that answering the additional recall module does not affect the accuracy of diaries: households randomized to, and households excluded from, the recall module have the same distribution of acquisitions in diaries. Besides, important departures from rank invariance are found across the support of diary and recall measurements and of the income distribution, suggesting that survey effects are unlikely to arise only due to households with specific demographics and income. These results imply that the same household could end up above or below the poverty line depending on the survey mode employed, which has distribution-wide implications for welfare ranking relevant to policies using proxy means tests derived from consumption data. The departure from rank invariance is also worrisome because it is among the

[3]These assumptions are rejected in our data and would be likely violated in most empirical settings (as conjectured by Aguiar and Bils, 2015).

[4]When the between-group component of inequality is of interest, however, the same argument can be used to claim that data from diaries dominate recall measurements.

weakest assumptions needed to address measurement error in empirical work.

Our research design yields a second important finding: the presumption that diaries should be the benchmark for measuring household well-being is an illusion. Although a consumer demand model implies that acquisition is an unbiased estimate of usual consumption, we demonstrate empirically that the difference between these two quantities can be substantial. For example, the likelihood of measuring less than half of the actual value of usual consumption with a diary is 15.1% and 14.1% for households in the first and fourth consumption quartiles, respectively. The likelihood of attributing a value for usual consumption at least twice the size of the actual one is 4.1% and 3.7% for these households, respectively. Importantly, we find that diary errors remain large for perishable components of food, which implies that our conclusions are not mechanically driven by consumption from stocks or large cash outlays. We conclude that, if enumerator visits are effective to reduce misreporting in diary entries, mistaken conclusions about welfare measurement from diary data must primarily depend on how households decide to smooth their consumption rather than on the frequency of acquisitions. This implies that even when error-free measurements of consumption in the interview week are attained, misleading conclusions about actual well-being could still ensue, as well-being is a longer-run average of weekly intakes.

Our investigation does not find evidence of zero-mean or classical errors in recall data, which is a conclusion often conjectured in the literature. Our research design allows us to prove this conjecture, and also provides guidance on how to sign the bias from recall errors in empirical work. Specifically, we show that – in our case study – overreporting of usual consumption is between three and four times more likely than underreporting. This finding is not only consistent with telescoping effects but also with more general cognitive errors by respondents when they compute the market value of their consumption. Moreover, we find that the quality of recall data improves with the actual value of usual consumption, which suggests that errors may be comparatively less important for households with higher permanent income and human capital.

Further, the above properties of survey errors have important consequences for the computation of aggregate welfare statistics in our study. Precisely, we find that diary errors yield poverty rates and inequality figures that are less reliable than those obtained from recall data, which yield uniformly larger probabilities of correct classification for households with low usual consumption. However, we also find that the upper tail of the distribution of usual consumption is closer to that from diary data, suggesting that deciding which survey measurement will work best ultimately depends on the policy question of interest.

What are the implications for the design of household surveys? First, there is no loss in accuracy from using a recall module despite the higher costs of using a diary: a recall survey is almost as close to the optimal design as a diary when inequality measurement is of concern. The choice of the most appropriate collection mode can be seen as the solution to a decision problem in which each household is assigned to a diary or a recall module to minimize the overall impact of survey errors. We therefore study which assignment would make the inequality and poverty measurements computed from raw data as close as possible to their true values. We show that the optimal solution to this problem is a mix of diaries and recall interviews, with relatively more well-off households assigned to diaries. Although a

survey employing only recall interviews is suboptimal in our case study, we find no evidence that such survey would compromise the accuracy of policy conclusions about poverty and inequality compared to a diary survey. We also show that recall distributions are closer to the underlying distribution of usual consumption than in diaries. Thus, our findings provide the theoretical and empirical rationale for the choice of switching to recall modules in expenditure surveys.

Second, our empirical investigation demonstrates the cost-effectiveness of eliciting *repeated measurements* from the same household, which is in the spirit of Browning and Crossley (2009). More specifically, when diaries are the collection mode of choice – as in the Iraq IHSES – the inclusion of one additional module on recalled consumption for a subsample of interviewees will allow for learning about survey errors and disparities in well-being at relatively low cost. When a recall survey is chosen instead, resources should be allocated by statistical agencies to conduct follow-up diary interviews on a smaller sample of interviewees. Our methodology therefore offers actionable recommendations to design a protocol for harmonizing time series of poverty and inequality in those countries that consider a transition to new modes of data collection.

The remainder of the paper is organized as follows. The next section presents the general formulation of the problem. Section 3 describes the institutional background and presents the diary-recall experiment. Section 4 presents the conditions for nonparametric identification. Estimation is also discussed, along with possible threats to the validity in our research design. Section 5 presents our results on the nature and consequences of survey errors. Section 6 looks at the optimal combination of diary and recall interviews. Section 7 concludes the paper.

2 General Formulation of the Problem

2.1 Quantity of Interest

Suppose that the household has a target monetary amount of consumption over a period of τ time units. *Consumption* is equal to *acquisition* over this period. Without loss of generality the time unit coincides with the interview period, which is one *week* in our case study.

We define Y^* as the average household consumption *per time unit*, although we will often use the terms *usual consumption* or *consumption rate* as shorthand. The quantity Y^* is *not observable*, as it results from a utility maximization problem solved by the household. It is interpreted as the average weekly rate of consumption that one would measure only by observing the household over τ weeks.[5] The empirical challenge is to learn about the distribution of Y^*, $F_{Y^*}[y]$, using diary (Y^d) and recall (Y^r) measurements for the interview period. These are possibly inaccurate indicators of Y^*, and this problem is inherently related to the short-term nature of interviews. Even perfectly measured values of consumption in a random week (obtained, for example, from stock inflows and outflows) will be different from Y^* in general.

The distribution $F_{Y^*}[y]$ is needed to retrieve key inequality functionals for policy analyses, which are

[5]For example, if the household targets \$1000 of consumption over $\tau = 4$ weeks and the desired amount is consumed in this period, the usual rate of consumption in one week must be $Y^* = \$1000/4 = \$250$. In this example, 4 weeks and \$1000 are choices made by the household to maintain smooth living standards depending on preferences and material resources.

the quantities of interest here. Household and area characteristics are also available, the conditioning on which is left implicit in what follows.

2.2 Diary Interviews

Diaries are a self-administered form of data collection in which respondents must record their market purchases and the estimated market values of acquisitions from non-market sources. In our case study, the quantity Y^d is defined as acquisitions captured over the one week diary.

The theoretical case for differences between consumption Y^* and acquisition Y^d can be made with the aid of a model of frequency of purchase and consumer demand (see Meghir and Robin, 1992, and Coibion et al., 2020, for examples). Using the notation above, households decide simultaneously on the desired consumption over τ weeks (e.g., one month) and a purchasing strategy. Assume, for simplicity, that acquisition is only conducted through cash outlays. Let N^* be the desired number of purchases in one week implied by utility maximization. An implication of this choice is that Y^*/N^* must be the expected amount of each purchase. If N denotes the number of purchases recorded in the diary week, the assumption of purchases of a constant amount yields:[6]

$$Y^d = N\frac{Y^*}{N^*}. \tag{1}$$

The main take-away message from this model is that household acquisition in the diary week is a random variable centered at the quantity of interest Y^*. Specifically, in a randomly chosen diary week, we have $E[N|N^* = n^*, Y^* = y^*] = n^*$ and:

$$E[Y^d|Y^* = y^*] = y^*. \tag{2}$$

The model also implies that distributional indicators obtained from diary acquisitions may be liable to provide incorrect policy conclusions. For example, calculations reported in the Online Appendix show that if acquisitions occur independently at the same rate N^* across households, we have:

$$\frac{Var[Y^d]}{Var[Y^*]} = 1 + \frac{1}{N^*} + \frac{1}{CV(Y^*)^2 N^*},$$

where $CV(Y^*)$ is the coefficient of variation of consumption. The last expression implies that for items usually acquired once a week ($N^* = 1$), the variance of acquisition must be at least twice as large as the variance of consumption. When the coefficient of variation is equal to one, the variance of acquisition becomes three times the variance of consumption.

We conclude that acquisition diaries may yield a distorted idea of consumption Y^* even if all entries are recorded correctly. However, entries may not be exempt from errors. As the effects of fatigue on accuracy have long been known to researchers (Silberstein and Scott, 2004), it is common practice to assist households through frequent visits by enumerators during the recording process. It is generally presumed that this design yields the most accurate data (see Brzozowski et al., 2017, for discussion).

[6]The model can be easily generalized to allow for random spending around Y^*/N^*.

Misreporting of diary entries affects equation (1) by introducing a disturbance term on the right-hand side.

2.3 Recall Interviews

Many household surveys implemented in the developing world ask respondents to recall quantities in a preselected list of food items consumed over a fixed period and their associated market value. In our case study, the recall period is seven days, ending with the interview (Lyberg and Kasprzyk, 2004, Beegle et al., 2012, and Crossley and Winter, 2014 discuss the effects of changing this time window). Information is collected in two steps. The respondent is first asked if the items in the list were consumed during the past week, going vertically down in the survey form from the first item to the last. For the items for which there is a positive response, the quantity consumed and the unit of measurement are reported. The respondent then must backcast the monetary value of this quantity by reporting estimated expenditures (if market purchased) or market values (if self-produced, received as a gift, or received in barter). The monetary equivalent of recalled consumption (quantity consumed times implicit price per unit, or unit value) is the quantity we consider for the measurement Y^r.

Why should one expect differences between Y^* and recalled consumption Y^r? As discussed above, consumption in the week before the interview need not coincide with usual consumption. Arguments can be made in favor of using a longer reference period for recall questions, or asking directly about "usual week" consumption. However, limited cognitive abilities and the difficulty of recalling the timing of events may challenge respondents' ability to perform the computations. Estimating the monetary equivalent of the quantities consumed from non-market sources adds an additional layer of difficulty. Errors may arise because the respondent solves a prediction problem rather than reporting a noisy measurement. Households may form their answers using the information available, for example including other features of the survey, so that the value Y^r is their best predictor of the unobserved value Y^*. Experimental evidence in psychology research supports this interpretation (see Comerford et al., 2009). Rounding errors in providing the quantities consumed or telescoping (the act of recalling consumption occurring over a longer period of time as within the reference period; Neter and Waksberg, 1964) were found to be important factors in the computation. These concerns may be particularly salient in a context such as Iraq, where many households have relatively low educational attainment.

The textbook assumption of classical errors is violated in general, with no obvious direction of bias. For example, if the best prediction is obtained by respondents in terms of quadratic loss, recall errors must be centered at zero and not correlated with Y^r but correlated with consumption Y^*. If an absolute value loss is used instead, the household would report the median of Y^* given the information available (see Hyslop and Imbens, 2001, and Hoderlein and Winter, 2010).

2.4 Aggregation over Items

Our empirical analysis uses measurements computed by aggregating over a number of food items. The definitions above can be adjusted to allow for I items:

$$\begin{aligned} Y^d &= \sum_i Y_i^d = Y^* \left(\sum_i b_i^* \frac{N_i}{N_i^*} \right), \\ Y^r &= \sum_i Y_i^r, \end{aligned} \tag{3}$$

where Y_i^* is usual consumption of item $i = 1, \ldots, I$, $Y^* \equiv \sum_i Y_i^*$ is (total) food consumption, and $b_i^* = \frac{Y_i^*}{Y^*}$ is the budget share allocated to the item.

Even after aggregation, acquisition from diaries remains an unbiased estimate of usual consumption Y^*. Specifically, a generalization of the frequency of purchase model to the case of a vector of items still yields equation (2) for total food acquisition Y^d, as we show in the Online Appendix. We also show there that, despite the aggregation across items, severely misleading conclusions about inequality from diary data are still possible. For example, assuming that acquisitions are independent across goods and occur at the same rates N_i^*'s for all households yields:

$$\frac{Var[Y^d]}{Var[Y^*]} \geq 1 + \sum_{i=1}^{I} \frac{1}{N_i^*} \frac{Var(Y_i^*)}{Var(Y^*)}.$$

The lower bound can be significantly larger than one depending on assumptions about the correlation of consumption across items. For example, if consumption is independent across items and goods are acquired once a week ($N_i^* = 1$), the variance of Y^d must be at least twice as large as the variance of Y^*.

Equation (2) may still be valid if survey participation affects the purchasing behavior of respondents, or with misreporting. For example, recorded cash outlays tend to be higher at the beginning of the diary period (in the first day in particular; see Lyberg and Kasprzyk, 2004, and Silberstein and Scott, 2004). If this pattern arises because of antedated expenditures that would normally take place later in the same week, equation (2) is still valid. Moreover, acquisition Y^d is still an unbiased estimate of Y^* under more general forms of manipulation or misreporting (e.g., overspending on goods linked to social status or underreporting due to an embarrassment motive) that result in any zero-mean, possibly non-classical, disturbance term at the right hand side of Y^d in (3).

3 Data, Descriptives and Graphical Analysis

3.1 The Iraq Household and Socio-Economic Expenditure Survey

Iraq is an upper middle income, resource-rich, fragile, and conflict-affected country facing development challenges in line with those of far poorer countries. The median education level in 2012 was primary schooling. While infant mortality rates remain below the norm for similar countries, a third of Iraqi children aged 0-5 are stunted, 58 percent of adult males of working age are employed, and only one in

five women of working age participate in the labor market. International standards for measuring poverty and socioeconomic indicators were adopted after the fall of Saddam Hussein.

The implementation of a new LSMS began in 2007, the IHSES. We use the second round of this survey, which was carried out in 2012 to obtain official statistics on poverty, food security, income, labor market outcomes, health, and education. The survey is the basis for information on the household consumption component in the national accounts and for consumer price indices. The IHSES collects a more detailed labor and income module than the norm (relative to, for instance, countries where the LSMS is complemented by a Labor Force Survey) and has special modules to fill critical data gaps (such as anthropometrics and time use).

IHSES 2012 had an intended sample size of 25,488 households and a final sample of 24,750 households. The population was stratified in 119 districts (*qadahs*). Within each district, 24 census enumeration areas (EAs) were randomly selected, and 9 households were sampled within each EA. Teams of interviewers were responsible for fieldwork in two districts distributed over one year. In each month, interviews were conducted by each team in four randomly selected EAs, two from each of the assigned districts. It follows that, by design, households in one EA had an equal probability of being interviewed in any of the survey months.[7]

3.2 Food Measurements

A 7-day diary of the standard acquisition type was used in all interviews, with one respondent (the head of household or the most informed person) recording acquisitions on behalf of the household. The completion of the diary was assisted by enumerators, who visited households on alternate days to check the quality of entries, clarify any questions, and enter recorded data into computers. Instructions emphasized that households should record daily acquisitions of food and non-food commodities and meals away from home starting from the day after the first visit. Qualitative analyses confirmed a high level of enumerator effort expended to ensure that diary entries were correctly and regularly recorded.

One-third of households within each EA were also administered a recall module on food consumption during the first visit. Specifically, *two* (diary and recall) measurements are available for household 3, 6 and 9 in each EA, while *one* (diary) measurement is available for all remaining households. The availability of repeated measurements and the scale of the experiment mark a departure from past research on consumption measurement in developing economies. The recall module asked the market value of the quantities consumed by the household in the 7 days prior to the enumerator's visit. A list of expenditure groups was selected based on an assessment of their importance in IHSES 2007 diaries. Fieldwork and resource constraints meant that this list was consolidated to include 20 such groups of aggregated food items (we list these groups in Appendix Figure A.2).[8]

[7] One of the important rationales for year-long surveys is the ability to capture seasonality in consumption and livelihoods. Seasonality in consumption is experienced due to differences in the availability of food items and, importantly, due to the month-long observations of Ramadan, and two Eid holidays. To ease fieldwork, a team would visit the 2 EAs from one of the districts in wave 1 (days 1 to 14 of the month) and the 2 EAs from the other district in wave 2 (days 15 to 29) of the month. Moreover, within each wave, the teams alternated the EA visited each day.

[8] The scale of IHSES precluded the use of individual diaries, which is deemed the most reliable instrument in Beegle et al. (2012). Our experiment is interesting because the cheaper option is randomized (recall questions) while the expensive option (diaries) remains the baseline for all households surveyed. Our recall module bears the closest resemblance to the subset list

Randomization to different interview modes was adopted to guide an eventual transition to recall as a cheaper way of collecting IHSES data on food.[9] As non-food expenditures are already measured using different recall periods depending on the frequency of purchase, the experiment was limited to food measurements. Recall groups were defined and matched to their counterparts in the diary using the COICOP classification, yielding the quantities Y^d and Y^r defined above. Both quantities include household estimations of the market value of food items consumed that are not acquired from the market such as self-produced items and gifts. In practice, however, market purchases account for almost all entries in our case study. We used the square root of family size to equivalize survey measurements in our analysis. Acquisitions and valuations of recalled consumption, in Iraqi dinars, were transformed to US dollars throughout.

There are no incentives for households to report specific consumption levels, for instance, to meet thresholds for social programs. The most important social safety net, the Public Distribution System (PDS), is a universal in-kind food subsidy, and other much smaller public transfers were not means-targeted at the time of the survey.[10] A separate IHSES module collected information about ration items received, consumed, bartered, sold or given away by the household during the last 30 days. The diary recorded market transactions to purchase ration items over and above the monthly allocation, but these transactions in our data are rare and small in magnitude. Because our focus is on comparing diary entries with recalled consumption, the consumption of rationed products is excluded from our analysis.

3.3 Covariate Balance and Raw Distributions

Table 1 documents the demographic balance for households assigned to recall modules. This randomly selected group of roughly 8,000 households constitutes our "treatment" sample, while all remaining households form the "baseline" sample. This table shows treatment-control differences from regressions that control for strata (EAs) used in the randomization design. All variables are well balanced between groups, as can be seen from the small and nonsignificant coefficient estimates in the two panels.

Randomization to the recall module did not affect the reporting of diary entries, as shown in Table 2. Respondents might adjust their diary to reconcile aggregates with those in the recall module, as in bounded interviews (see Neter and Waksberg, 1964), but the results in column 2 rule out this possibility. Specifically, Panel A reports treatment-control differences in cumulative distributions of diary acquisitions Y^d at selected percentiles, as well as differences in averages. Treatment-control differences in the dispersion of Y^d are reported in Panel B. Results in column 2 imply that diaries collected in treatment and baseline samples yielded statistically indistinguishable distributions of household acquisitions.

Recall data yielded higher consumption figures than diary acquisition. This can be seen from column 3 in Panel A of Table 2, which presents coefficients from the same regressions in column 2 but using

of the 17 most important food items considered in the Beegle et al. (2012) Tanzania experiment.

[9] As noted in the sampling and fieldwork documentation, the recall module "*should be administered in the first visit to the household, before the recording of food consumption by diaries. Asking these questions afterwards (when both the respondent and the interviewer will know the diary records) would defeat the purpose of this module, which is to compare the results obtained from the two instruments, to assess the possibility of applying in future surveys the recall method instead of diaries*".

[10] The PDS includes 13 rationed products, of which four, whole wheat flour, rice, vegetable oil/cooking oil and sugar represent almost 98% of total rationed expenditures. Because they are heavily subsidized, ration expenditures account for only 6% of average household spending.

Table 1: Summary statistics and covariate balance[a]

	All Households		Rural Households		Urban Households	
	Baseline mean	Treatment difference	Baseline mean	Treatment difference	Baseline mean	Treatment difference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Household demographics						
Log income	6.1908	-0.0042	5.9676	-0.0265	6.2743	0.0041
	[*0.0079*]	(*0.0118*)	[*0.0114*]	(*0.0187*)	[*0.0097*]	(*0.0146*)
Log implicit rent	4.6725	-0.0091	3.8909	-0.0016	5.0159	-0.0126
	[*0.0120*]	(*0.0100*)	[*0.0133*]	(*0.0137*)	[*0.0124*]	(*0.0133*)
Log price index	0.9921	-0.0027	0.9489	-0.0019	1.0082	-0.0030
	[*0.0014*]	(*0.0018*)	[*0.0017*]	(*0.0023*)	[*0.0017*]	(*0.0024*)
Household size	6.7184	0.0843	7.7662	0.1452	6.3263	0.0614
	[*0.0365*]	(*0.0654*)	[*0.0609*]	(*0.1168*)	[*0.0432*]	(*0.0786*)
N. of children	2.1690	-0.0369	2.8358	0.0190	1.9195	-0.0580
	[*0.0228*]	(*0.0392*)	[*0.0385*]	(*0.0725*)	[*0.0269*]	(*0.0466*)
N. of adults	3.9653	0.0455	4.1670	0.0440	3.8899	0.0461
	[*0.0250*]	(*0.0450*)	[*0.0409*]	(*0.0730*)	[*0.0306*]	(*0.0556*)
N. of occupied adults	1.5105	-0.0076	1.5842	-0.0377	1.4829	0.0037
	[*0.0123*]	(*0.0212*)	[*0.0200*]	(*0.0354*)	[*0.0151*]	(*0.0259*)
Panel B. Household head demographics						
Male	0.8841	0.0005	0.9175	-0.0083	0.8716	0.0039
	[*0.0041*]	(*0.0069*)	[*0.0047*]	(*0.0090*)	[*0.0053*]	(*0.0089*)
Age	46.7432	0.1303	44.7398	-0.1067	47.4930	0.2194
	[*0.1661*]	(*0.2786*)	[*0.2388*]	(*0.4218*)	[*0.2087*]	(*0.3489*)
Has primary education	0.6419	-0.0183*	0.5626	-0.0132	0.6715	-0.0202
	[*0.0056*]	(*0.0099*)	[*0.0085*]	(*0.0153*)	[*0.0070*]	(*0.0124*)
Has secondary education	0.2527	-0.0093	0.1629	-0.0007	0.2863	-0.0125
	[*0.0056*]	(*0.0093*)	[*0.0067*]	(*0.0118*)	[*0.0072*]	(*0.0120*)
Reads and writes	0.7780	-0.0013	0.7161	-0.0022	0.8012	-0.0010
	[*0.0047*]	(*0.0081*)	[*0.0076*]	(*0.0122*)	[*0.0058*]	(*0.0101*)
Employed	0.7313	-0.0108	0.7514	-0.0164	0.7238	-0.0087
	[*0.0054*]	(*0.0097*)	[*0.0075*]	(*0.0140*)	[*0.0069*]	(*0.0122*)
Number of districts	119	119	119	119	119	119
Number of EAs	2,828	2,828	2,828	2,828	2,828	2,828
Number of households	16,530	8,220	16,530	8,220	16,530	8,220

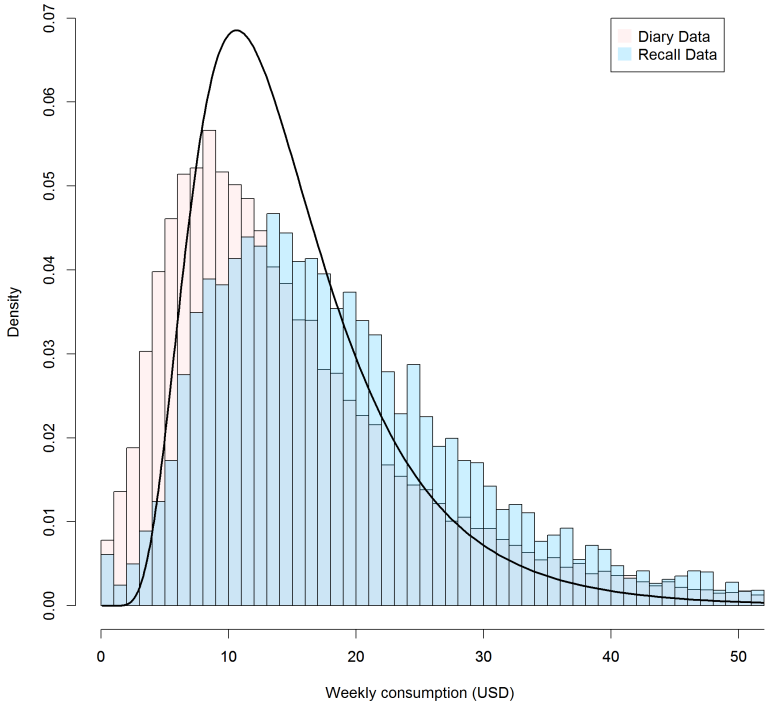
[a] This table reports tests for covariate balance using household-level demographics (Panel A) and demographics of the household head (Panel B). Household is the unit of analysis. Columns (1), (3) and (5) report statistics for the *baseline sample*, which consists of households excluded from the recall-module experiment. Means and standard deviations (in square brackets) are reported for variables listed at left. The *treatment sample* consists of households randomized to the recall-module experiment. For these households, both diary and recall measurements are available. Columns (2), (4) and (6) report coefficients from regressions of each variable on the treatment dummy and a full set of dummies for strata (enumeration areas, EAs) used in the randomization design, pooling data from the two samples. Standard errors for the coefficient on the treatment dummy (in round brackets) are clustered on EAs. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Between-sample differences in survey measurements[a]

	Baseline	**Treatment Difference**	
	Diary	Diary-Diary	Recall-Diary
	(1)	(2)	(3)
		Panel A. Cumulative Distribution and Location	
5th percentile	2.5291	0.0059	-0.0431***
		(0.0045)	*(0.0042)*
25th percentile	2.0622	0.0121	-0.1482***
		(0.0085)	*(0.0086)*
50th percentile	2.5291	0.0097	-0.1731***
		(0.0099)	*(0.0101)*
75th percentile	2.9788	0.0100	-0.0944***
		(0.0085)	*(0.0093)*
95th percentile	3.6601	0.0032	-0.0032
		(0.0037)	*(0.0039)*
Mean	2.5184	-0.0158	0.2706***
		(0.0099)	*(0.0105)*
		Panel B. Dispersion	
Std. deviation	0.7560	0.0182	-0.0650***
		(0.0157)	*(0.0244)*
Range 10%-90%	1.8207	0.0044	-0.3227***
		(0.0387)	*(0.0320)*
Range 25%-75%	0.9166	-0.0044	-0.1286***
		(0.0236)	*(0.0207)*
Number of districts	119	119	119
Number of EAs	2,828	2,828	2,828
Number of households	16,530	24,750	24,750

[a] Household is the unit of analysis. Column (1) reports selected statistics of the $\log(Y^d)$ distribution in the *baseline sample*. Panel A: Columns (2) and (3) report coefficients on the treatment dummy in regressions that control for a full set of strata effects (enumeration areas, EAs) pooling data for treatment and control samples. Standard errors (in brackets) are clustered on EAs. For cumulative distributions, the outcome in column (2) is an indicator for having $\log(Y^d)$ below the value in column (1); the outcome in column (3) is an indicator for having $\log(Y^d)$ below the value in column (1) for households in the baseline sample, and an indicator for having $\log(Y^r)$ below the value in column (1) for households in the treatment sample. For the mean, the outcome in column (2) is $\log(Y^d)$. The outcome in column (3) is $\log(Y^d)$ for households in the baseline sample, and $\log(Y^r)$ for households in the treatment sample. Panel B: The value in column (2) is the difference of statistics computed using $\log(Y^d)$; the value in column (3) is the difference of statistics computed using $\log(Y^d)$ for households in the baseline sample, and $\log(Y^r)$ for households in the treatment sample. Standard errors for these differences are computed via bootstrap using 1,000 replications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1: Survey measurements and usual consumption

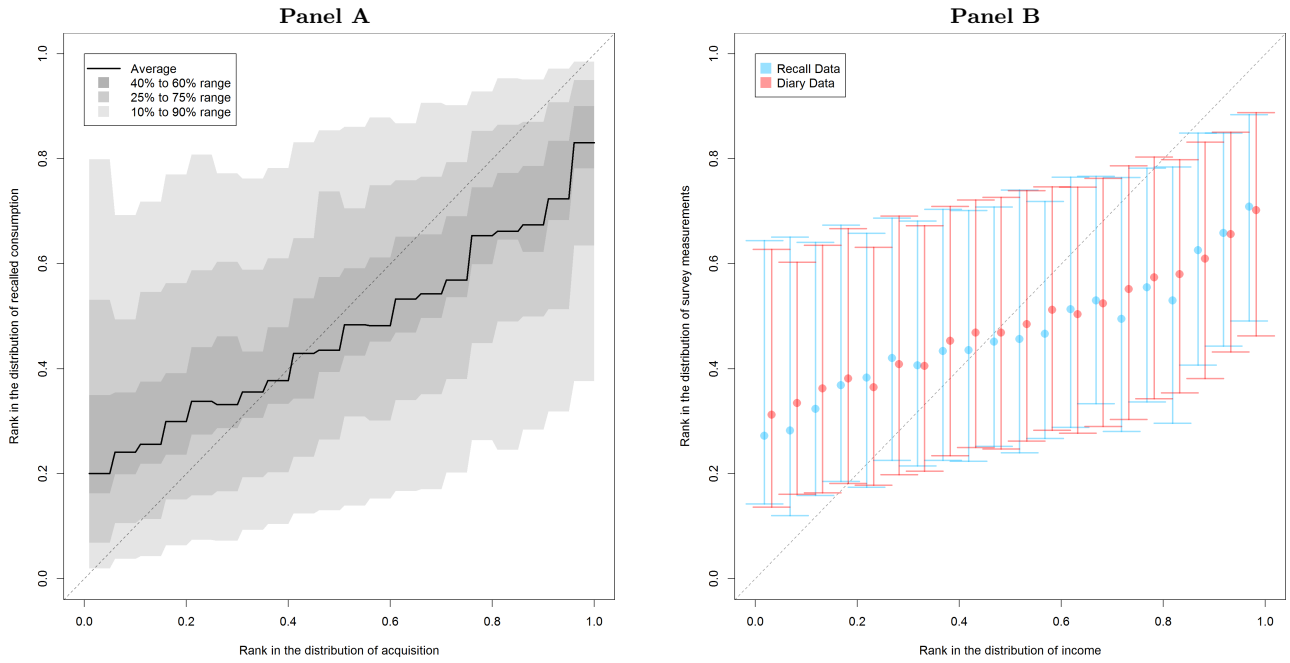


Note. This figure shows the empirical densities of food measurements (histograms) and the estimated density of usual consumption (continuous line). Household is the unit of observation. Acquisitions and valuations of recalled consumption, in Iraqi dinars, are reported in US dollars (USD) on the horizontal axis. The histogram for diary is computed using all households, while the histogram for recall is computed from the *treatment sample* – see note to Table 1 for definitions.

diary reports Y^d for the baseline sample and recall reports Y^r for the treatment sample. Larger recalled consumption may be a counterintuitive result, as diary surveys should have less memory loss. However, this result is consistent with findings from other contexts, including Canada (Brzozowski et al., 2017), Niger (Backiny-Yetna et al., 2017), Tanzania (Beegle et al., 2012), and the United States (Battistin, 2003, and Bee et al., 2013). This result is also consistent with smaller diary-recall differences for recall periods longer than the one considered here, a fact first documented in work by Neter and Waksberg (1964): the memory of respondents declines with the length of the recall period, leading to lower recall aggregates. Another possible explanation is the telescoping of consumption.

A visual illustration of diary-recall differences across the support of the consumption distribution is in Figure 1. The mode of interview effects varies across households: differences between distributions are not a simple location shift and indicate different effects across percentiles. This evidence weighs against mean diary-recall differences as the most interesting quantity to consider and suggests that the interview mode must be accompanied by subtly nuanced effects on measured inequality that cannot be understood by a simple comparison of means. In addition, column 3 in Panel B of Table 2 shows that the consumption distribution from recall data has lower variance than the one from diaries.

Figure 2: Household rank in diary and recall distributions



Note. Panel A compares diary and recall reports from the same household. Rank invariance corresponds to the dashed line (the 45-degree line). The solid line represents the average percentile in the recall distribution for households in the p-th percentile of the diary distribution. The darker areas represent the $40\% - 60\%$, $25\% - 75\%$ and $10\% - 90\%$ ranges for these households. Panel B compares diary and recall reports with household income. Dots are average percentiles in the diary or recall distributions for households in the p-th percentile of the income distribution. The figure also reports the $25\% - 75\%$ range for these households.

3.4 Ranks

Diary and recall reports from the same respondent do not order her household identically in the population, as shown in Panel A of Figure 2. Here, the horizontal axis shows household ranks computed from diaries Y^d. On the vertical axis, the same households are ranked in the recall distribution Y^r. The analysis is carried out using the sample for which both measurements are available. The continuous line in the figure shows the average percentile in the recall distribution for households sharing the same percentile in the diary distribution. Shaded areas are obtained in a similar manner by considering percentile ranges instead of the average percentile. The two measurements are clearly not rank preserving. This finding is consistent with prior work by Battistin and Padula (2016), who reached the same conclusion on data for the United States.[11]

Differences between recall and diary measurements are not mechanically explained by household income. This can be seen in Panel B of Figure 2, where diary and recall ranks on the vertical axis are plotted against ranks in the household income distribution. Consumption measurements flatten the difference in well-being across households depicted by income. For example, households in the bottom quintile of the

[11]The frequency of purchases is the most likely explanation for these findings. Calculations reported in the Appendix show that households are between two and three times more likely to recall consumption of storable items than to report acquisition of these items in the diary week. For example, 40% of households report positive acquisition of rice in the diary week, while almost all households recall a positive consumption of rice in the week preceding the diary. The frequency effects on survey reports apply to respondents from all socioeconomic backgrounds. Differences between percent acquiring and percent recalling consumption shrink towards zero for more perishable items, such as meat or fish (see Appendix Figure A.2).

income distribution are, on average, in the second quintile of the consumption distribution. Households in the top income quintile are, on average, in the third consumption quintile. This relationship does not change with the survey mode employed, suggesting that diary-recall differences cannot be explained by differential errors at different points of the income distribution.

4 Nonparametric Identification

4.1 Assumptions

We assume that usual consumption Y^* and its measurements (Y^d, Y^r) are continuously distributed with bounded density. Consumption is defined by aggregation over items. The distribution of consumption on each item need not be continuous, for example because of zero consumption on single items. The assumption here is that aggregation across items makes the distribution of household consumption smooth enough. A visual inspection of the data corroborates this idea. Moreover, the theoretical case for lognormally distributed consumption is made in Battistin et al. (2009).

We maintain the assumption that the difference between household consumption and acquisition is centered at zero. The theoretical case for this restriction stems from the consumer demand model leading to Y^d in equation (3) – see also the Online Appendix. As we discussed above, the following assumption also allows for possibly nonclassical reporting errors in diary entries.

Assumption 1 *(Unbiasedness of Diary Measurements)* $E[Y^d - Y^*|Y^* = y^*] = 0$.

Nonparametric identification requires additional restrictions imposing structure between observables and usual consumption that we discuss with the aid of a directed acyclic graph (DAG). In Figure 3 survey measurements Y^d and Y^r depend on consumption Y^*. We require the existence of an observable variable Z which affects survey measurements only through consumption.

Assumption 2 *(Exclusion Restriction) There exists a continuous variable* Z *such that* $(Y^r, Y^d) \perp Z|Y^*$.

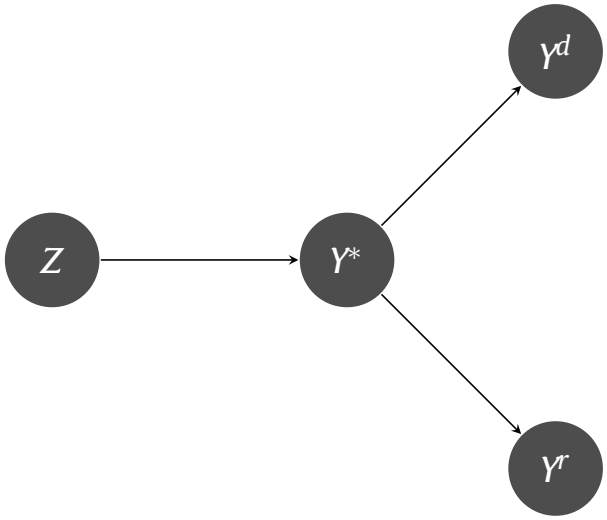
This is a standard exclusion restriction in the literature which has been used and justified in a number of empirical settings, for example, in education, labor and earning dynamics (see Schennach, 2013, and Wilhelm, 2019). It can be interpreted through the lens of the (unfeasible) regressions of Y^r and Y^d on Y^*: the variable Z must be an excluded instrument for Y^* in both equations. A more general interpretation follows from the requirement of conditional independence: knowledge of Z must not yield any more information on the reporting of survey measurements than Y^* would otherwise provide.[12]

The DAG in Figure 3 also imposes that Y^* is the only common cause for the correlation among survey reports, which is the meaning of Assumption 3 below.

Assumption 3 *(Conditional Independence) Survey measurements are conditionally independent given usual consumption* $Y^r \perp Y^d|Y^*$.

[12]This assumption is also known as *surrogacy*, meaning that the surrogate variable Y^* must mediate the link between Z and the survey reports (see, for example, Athey et al., 2019).

Figure 3: Directed acyclic graph of model assumptions



Note. A directed acyclic graph (DAG) is a collection of nodes representing random variables and directed edges among nodes representing causal effects which are not mediated by other variables. This figure is a visual representation of the key assumptions used to achieve non-parametric identification. Diary and recall measurements from the same household are denoted by Y^d and Y^r, respectively. These survey measurements are manifestations of usual consumption Y^*, which is unobserved. The theoretical case for having survey measurements differ from consumption Y^* is made in Section 2. Assumption 2 says that the *observable* variable Z affects *unobservable* consumption Y^* and is correlated with survey measurements only through Y^*. Assumption 3 implies that the correlation between survey measurements Y^d and Y^r follows from the fact that they both measure Y^*. The model in Section 2 sheds light on the meaning of this assumption.

This assumption is sufficiently general to allow for correlation between errors in recalled consumption and errors in diary acquisitions. Assumption 3 also allows for correlation between the respondents' ability to recall and their usual acquisition frequencies N_i^*, as can be seen from equation (3). In this respect, our analysis allows for forms of misreporting which are considerably more general than independent and classical survey errors. On the other hand, the assumption imposes that only *one* unobservable drives the correlation between survey measurements. For example, the equations in (3) show that this restriction would be violated if, after conditioning on Y^*, recall errors were still correlated with household preferences over frequency of acquisition.

Two additional formal conditions are needed for identification, which are implicit in Figure 3 and unlikely to be a limitation in our setting. These conditions are *not* testable because Y^* is unobservable. First, usual consumption must have a causal effect on the distribution of recalled consumption Y^r, as shown in Figure 3. The next condition is satisfied, for example, if $E[Y^r|Y^*]$ increases monotonically in Y^*, but it is much weaker than that.

Assumption 4 *(Informativeness of Recall) The relationship between Y^r and Y^* satisfies $F_{Y^r}[y|Y^* = y_1^*] \neq F_{Y^r}[y|Y^* = y_2^*]$, for any values $y_1^* \neq y_2^*$.*

We also require the following condition.

Assumption 5 *(Completeness) The relationship between Y^d, Y^* and Z satisfies:*

1. *Y^d is complete for Y^*;*

2. $Y^*$is complete for Z.

Completeness is common in the literature on nonparametric identification with instrumental variables (as in Newey and Powell, 2003), and is a weak regularity condition. Formally, completeness of Y^d for Y^* is guaranteed if $E[h(Y^d)|Y^* = y^*] = 0$ for all y^* only when $h(Y^d) = 0$. Intuitively, we require sufficient variability in the conditional distribution of Y^d at different values of Y^* and, similarly, for the conditional distribution of Y^* given Z. The latter requirement is a generalization of the rank condition for the instrument Z to the non-parametric setting.

4.2 Empirical Specifications

Assumptions 1-5 imply:

$$f_{Y^d Y^r|Z}[y^d, y^r|z] = \int f_{Y^d|Y^*}[y^d|y^*] f_{Y^r|Y^*}[y^r|y^*] f_{Y^*|Z}[y^*|z] dy^*, \tag{4}$$

and that there exists only one set of unobservable distributions on the right-hand side yielding the observable distribution on the left-hand side. In other words, under the assumptions stated, data on (Y^d, Y^r, Z) are sufficient to retrieve the distribution of usual consumption:

$$f_{Y^*}[y^*] = \int f_{Y^*|Z}[y^*|z] f_Z[z] dz,$$

and the distributions of diary survey errors ($f_{Y^d|Y^*}[y^d|y^*]$) and recall survey errors ($f_{Y^r|Y^*}[y^r|y^*]$). This identification result does not hinge on any parametric assumption, and follows from adapting results in Hu and Schennach (2008) to the case of multiple (diary and recall) survey measurements.

We constructed Z by matching each household to the *set* of all other households in the *same* income decile, filling out the diary in the *same* survey wave and receiving visits from a *different* team of interviewers. We defined Z as average diary acquisition in this set, which is also an estimate of the average of Y^* in this set because of Assumption 1. The precision of this estimate depends on the sample size used to compute Z for each household, which in our data is about 100 households. The idea then is to use the correlation between a household's own consumption, Y^*, and the average consumption of similar peers, Z, to learn about the properties of survey measurements. Considering peers in the same income decile and interviewed in the same wave strengthens the case for this correlation. The key exclusion restriction here is that variability in the average consumption of peers in different EAs (and thus visited by different interviewers) cannot explain a household's diary and recall reports other than through her Y^*. Violations of this assumption due to interviewer effects are ruled out by defining Z from averages of diary acquisitions supervised by non-overlapping survey teams.

We estimated the conditional densities on the right hand side of equation (4) using a flexible class of exponential families (Barron and Sheu, 1991). We let the density of $\log Y^r$ given $\log Y^*$ depend on a vector of parameters $\boldsymbol{\beta}^r$ as follows:

$$f_{\log Y^r|\log Y^*}[\log y^r|\log y^*; \boldsymbol{\beta}^r] = M(\boldsymbol{\theta}^r) \exp\left\{\sum_{k=1}^{K_r} \theta_k^r L_k\left(\frac{\log y^r - \log y^*}{\Delta_r}\right)\right\}, \tag{5}$$

where the normalizing constant is:

$$M(\boldsymbol{\theta}^r) = \left[\int \exp\left\{ \sum_{k=1}^{K_r} \theta_k^r L_k \left(\frac{\log y^r - \log y^*}{\Delta_r} \right) \right\} d \log y^r \right]^{-1},$$

$L_k(x)$ is the k-th Legendre polynomial defined on the interval $[-1, 1]$, and:

$$\theta_k^r = \sum_{j=0}^{J_r} \beta_{kj}^r L_j \left(\frac{\log y^* - \delta_0^r}{\delta_1^r} \right).$$

We set $\Delta_r = 4$ to ensure that the argument of $L_k(x)$ lies in the interval $(-1, 1)$ for any reasonable choice of $\log y^r$ and $\log y^*$. The smoothing constant K_r determines the degree of departure from normality of the distribution in (5). The value $K_r = 2$ is equivalent to imposing a normal distribution, while a value $K_r > 2$ allows for significant departures from normality. The quantities δ_0^r and δ_1^r were set to 3 and 7, respectively, to ensure that the argument of $L_j(x)$ is in $(-1, 1)$. Our empirical investigation considers $J_r = 2$, thus allowing each parameter θ_k^r to depend on $\log y^*$ via a quadratic function. This specification implies that the density in (5) depends on a vector of $3 \times K_r$ parameters $\boldsymbol{\beta}^r = (\beta_{10}^r, \beta_{11}^r, \beta_{12}^r, \ldots, \beta_{K_r 0}^r, \beta_{K_r 1}^r, \beta_{K_r 2}^r)'$.

We used the same specification for the density of $\log Y^d$ given $\log Y^*$, which depends on a $3 \times K_d$ vector $\boldsymbol{\beta}^d$ (the density-specific smoothing constant K_d is defined by analogy with K_r). Finally, we let the density of $\log Y^*$ given $\log Z$:

$$f_{\log Y^* | \log Z}[\log y^* | \log z; \boldsymbol{\beta}^y] = M(\boldsymbol{\theta}^y) \exp\left\{ \sum_{k=1}^{K_y} \theta_k^y L_k \left(\frac{\log y^* - \log z}{\Delta_y} \right) \right\},$$

depend on a $3 \times K_y$ vector $\boldsymbol{\beta}^y$, we set:

$$\theta_k^y = \sum_{j=0}^{J_y} \beta_{kj}^y L_j \left(\frac{\log z - \delta_0^y}{\delta_1^y} \right),$$

with $\Delta_y = 8$, $\delta_0^y = 0$, $\delta_1^y = 3$, and $J_y = 2$, and $M(\boldsymbol{\theta}^y)$ is the normalizing constant.

The unknown parameters $\boldsymbol{\beta}^r$, $\boldsymbol{\beta}^d$ and $\boldsymbol{\beta}^y$ were estimated by sieve maximum likelihood (Grenander, 1981), maximizing the pseudo-likelihood obtained by substituting the unknown conditional densities in (4) with their approximations described above.[13] By allowing the smoothing constants K_r, K_d and K_y to increase with sample size, this procedure yields non-parametric estimates of the conditional densities of interest. We ended up selecting $K_r = K_d = 6$ and $K_y = 3$, as negligible improvements to the likelihood were found for larger values of these parameters.

Both diary and recall error distributions are therefore estimated away from normality. The conditional distribution of logged consumption is also not normal, although the degree of departure from this

[13] Log transformations were considered to ensure numerical stability. Distributions of variables in levels, used in the sections below, can be obtained straightforwardly from the distribution of their logs. The likelihood was constructed using all households in the sample. Under Assumptions 1-5 the contribution to the likelihood of households not answering the recall module is:

$$\int f_{Y^d | Y^*}[y^d | y^*] f_{Y^* | Z}[y^* | z] dy^*.$$

Table 3: Survey measurements, usual consumption and survey errors[a]

	Measurements		Consumption	Survey Errors: Mean	Survey Errors: MSE	
	Diary	Recall		Recall	Diary	Recall
	(1)	(2)	(3)	(4)	(5)	(6)
1st decile	4.7270	7.4550	7.1882	1.408	0.288	0.620
2nd decile	6.8615	9.7729	8.8986	1.356	0.267	0.512
3rd decile	8.5786	11.9128	10.3768	1.320	0.257	0.446
4th decile	10.3563	14.1222	11.8958	1.290	0.250	0.396
5th decile	12.4364	16.2122	13.4062	1.265	0.246	0.358
6th decile	14.7070	18.9989	15.2380	1.239	0.243	0.321
7th decile	17.5878	21.9847	17.3200	1.214	0.242	0.288
8th decile	21.7675	26.0026	20.3705	1.185	0.242	0.253
9th decile	29.4806	32.9604	25.2176	1.149	0.246	0.214
Mean	15.2130	18.6712	15.2050			
Gini index	0.3765	0.3143	0.2656			

[a] This table compares distributions of survey reports and usual consumption. Columns (1) considers data on Y^d for all households. Columns (2) considers data on Y^r for the *treatment sample* – see note to Table 1 for definitions. Column (3) reports the estimated distribution of Y^*. Column (4) reports the estimated mean of Y^r/Y^* conditional on deciles of Y^* at left. The mean of Y^d/Y^* is not reported, and this is always one because of Assumption 1. Column (5) reports the Mean Squared Error (MSE) of Y^d/Y^* conditional on deciles of Y^* at left. Column (6) shows the MSE of Y^r/Y^* conditional on deciles of Y^* at left.

benchmark is less pronounced.

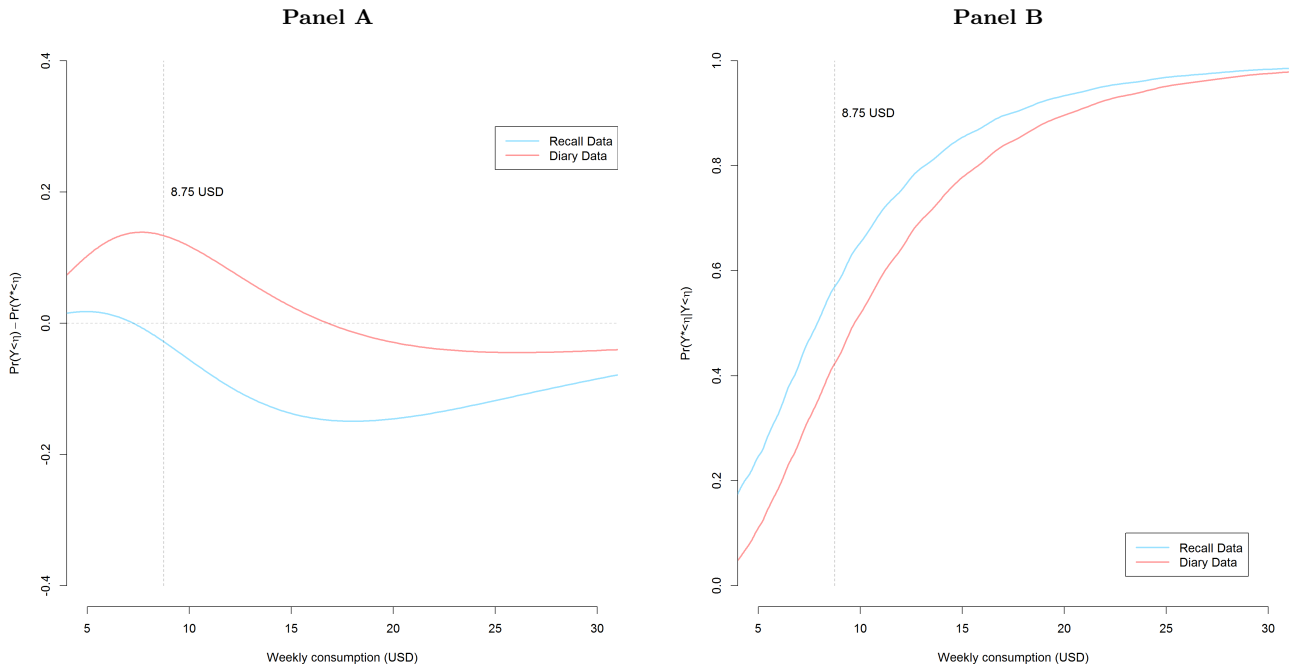
5 Errors in Survey Reports

5.1 Diary and Recall Measurements

The presumption that diaries outperform recall data for the measurement of well-being finds little empirical support. This can be seen in Figure 1, where diary and recall histograms are compared with the distribution of usual consumption, $f_{Y^*}[y]$ (the continuous line). The lower tail of the recall distribution is closer to that of usual consumption, and mismeasurement from diaries is more substantial at the bottom end. As diary measurements are centered at Y^* (because of Assumption 1), the difference in the modes of $f_{Y^*}[y]$ and $f_{Y^d}[y]$ suggests that diary errors must have a thick lower tail. Figure 1 also illustrates how the upper tail of the distribution of recall data is substantially thicker than the distribution of usual consumption, which is better approximated using data from diaries. These conclusions are confirmed by the first three columns of Table 3, where different percentiles of empirical and estimated distributions are compared. The bottom three deciles of the recall distribution are closer to the deciles of Y^* than are those of the diaries. The opposite conclusion applies for the remaining deciles. Consistent with predictions from the model in Section 2, the Gini coefficient from diaries overstates the real value of this index; the distance from the real value is minimized using recall data.

These findings have important implications for the computation of aggregate poverty statistics. This

Figure 4: Poverty mismeasurement and the misclassification of households

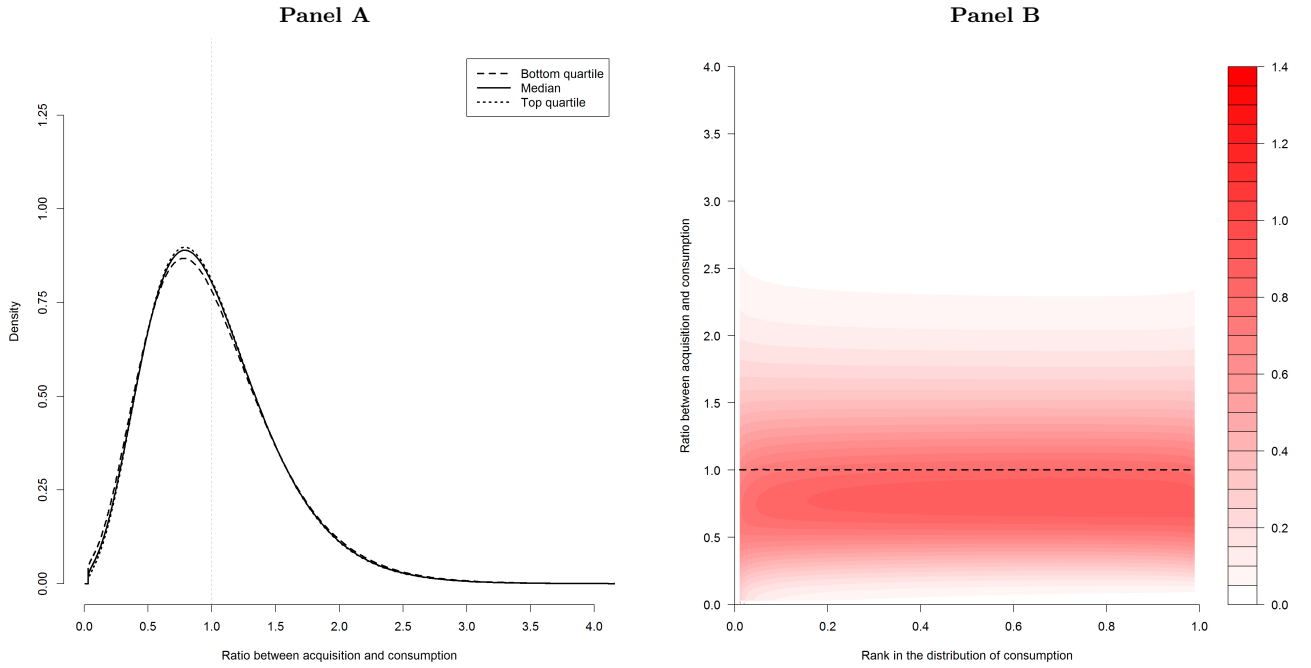


Note. The lines in Panel A show the difference between the share of households with measurements below η, $P\left(Y^d \leq \eta\right)$ or $P\left(Y^r \leq \eta\right)$, and the same quantity computed using the distribution of usual consumption, $P\left(Y^* \leq \eta\right)$. The lines in Panel B show $P\left(Y^* \leq \eta | Y^d \leq \eta\right)$ and $P\left(Y^* \leq \eta | Y^r \leq \eta\right)$, which are probabilities of correct classification in diary and recall measurements, respectively, at different values of η (which are in US dollars, USD).

can be seen in Panel A of Figure 4, where any point η on the horizontal axis is a value in the support of Y^*. The curve for diary here is $P\left(Y^d \leq \eta\right) - P\left(Y^* \leq \eta\right)$, which represents the error in the proportion of households with usual consumption below η. The curve for recall, $P\left(Y^r \leq \eta\right) - P\left(Y^* \leq \eta\right)$, has a similar interpretation. For example, with an absolute poverty line of 1.25 USD a day/person, the error is markedly higher using diaries: the share of households consuming less than $\eta = 8.75$ USD a week (the bottom 19% in the distribution of usual consumption) is 16.1% in the recall distribution and 32.3% using diaries. These differences yield the values 13.3% and -2.9% for the diary and recall curves in Panel A, respectively, when $\eta = 8.75$. Similar calculations – not shown here – can be used to compute the measurement effects on the poverty gap, which measures the average distance to the poverty line (where households above the line are given a distance of zero). The value of this index is 11% in diaries and 4.1% in recall data, which should be compared with the value of 2.2% obtained using $f_{Y^*}[y]$.

Means testing obtained from the diaries is not bullet proof either, as we show in Panel B of Figure 4. The curve shown for the diaries here is $P\left(Y^* \leq \eta | Y^d \leq \eta\right)$, which is the probability that a household with acquisitions lower than η also consumes less than η. For example, by setting $\eta = 8.75$ USD a week, the quantity $P\left(Y^* \leq \eta | Y^d \leq \eta\right)$ is the probability that households identified as poor according to the diaries are genuinely consumption poor. The curve $P\left(Y^* \leq \eta | Y^r \leq \eta\right)$ for recall can be interpreted similarly. Panel B of Figure 4 shows that the probability of correct classification for diaries is 42.3% when $\eta = 8.75$ USD a week. Although recall data yield uniformly larger probabilities of correct classification, a large amount of error still remains. For example, the probability of correct classification using recall is 57%

Figure 5: Survey errors in acquisition diaries



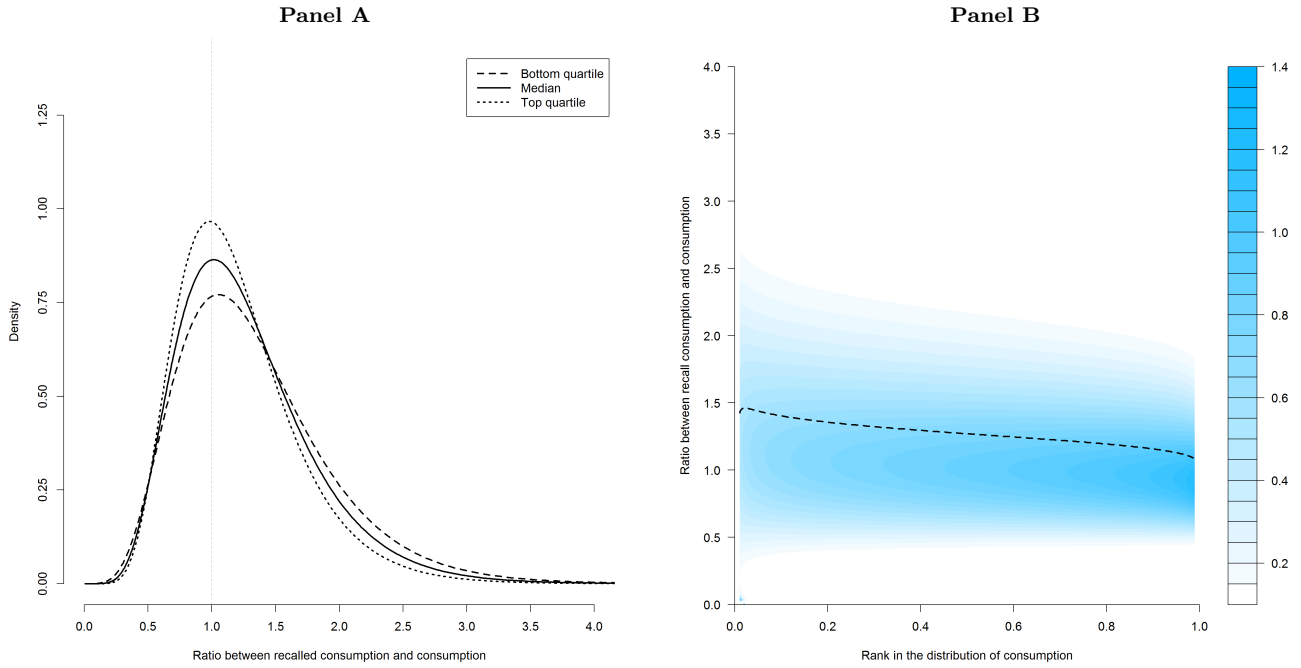
Note. Panel A shows error distributions for a hypothetical household at three selected percentiles of Y^*. The quantity Y^d/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from diaries, Y^d, and her usual consumption. No diary error implies a value of one for this ratio. Panel B shows a contour plot for the same distributions, where darker colors denote higher density. The dashed line here represents the conditional mean, which is equal to one by construction (see Assumption 1).

when $\eta = 8.75$ USD a week.

The assumption of error-free diaries is an illusion. This can be seen from Panel A of Figure 5, which reports error distributions for a hypothetical household at selected percentiles of Y^*. The quantity Y^d/Y^* on the horizontal axis lends itself to a simple interpretation: for instance, a value of 1.5 here means that the household measurement is 50% larger than its usual consumption. Because of Assumption 1, all distributions in this panel are centered at one. Despite being correct on average, we find that diary measurements most likely understate usual consumption as the mode of distributions is always below one. The chance of severely understating consumption ($Y^d/Y^* < 0.5$) is 15.1% and 14.1% for a household in the first and third quartiles, respectively. The chance of attributing a value to household consumption at least twice as large as the real one ($Y^d/Y^* > 2$) is 4.1% and 3.7% in these quartiles. Panel B of Figure 5 shows the family of densities Y^d/Y^* across all percentiles of Y^*. In this contour plot, quantiles are on the horizontal axis, darker colors denote a higher probability mass, and the dashed line shows the conditional expectations (which are equal to one). The mode of these distributions is always below one, meaning that underreporting is the most likely outcome. The contour plot also shows the large spread of error distributions.

Errors in recalled consumption are far from being classical in form, with overreporting being more likely than underreporting. This can be seen in Panel A of Figure 6, which reports error distributions considering the values of Y^r/Y^* on the horizontal axis. The contour plot in Panel B shows that the conditional means and modes become closer to one as usual consumption increases. For example, the

Figure 6: Survey errors in recalled consumption



Note. Panel A shows error distributions for a hypothetical household at three selected percentiles of Y^*. The quantity Y^r/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from the recall module, Y^r, and her usual consumption. No recall error implies a value of one for this ratio. Panel B shows a contour plot for the same distributions, where darker colors denote higher density. The dashed line here represents the conditional mean.

averages of the distribution Y^r/Y^* for households in the 10th and 90th percentiles of Y^* are 1.41 and 1.15, respectively, as shown in column (4) of Table 3. The average error for a household with median consumption is 1.27. Recall errors are therefore negatively correlated with consumption.[14] Moreover, the standard deviation of these distributions shrinks as consumption increases.

The simplest explanation seems most likely: higher usual consumption is correlated with human capital, better cognitive abilities, and the ability to compute more reliable recall measurements. Poorer households tend to overreport consumption. The likelihood of severely underreporting consumption using recall questions ($Y^r/Y^* < 0.5$) is 3.2% and 2.4% for households in the first and third quartile, respectively, while the probability of reporting a value for household consumption at least twice as large as the real one ($Y^r/Y^* > 2$) is 12.5% and 6.2% for respondents in the same two quartiles. The time pressure argument would suggest that those with higher incomes and less leisure should be less likely to respond accurately to surveys. We find the opposite pattern.

Given the trade-offs involved, determining which survey method works best ultimately depends on the policy question. Columns (5) and (6) of Table 3 show that diaries provide the most accurate predictor of household consumption in terms of mean squared error (MSE). It follows that, when predictions about individual households consumption are the quantity of interest, diaries should be used instead of recall modules. We note, however, that the mode of recall measurements is close to that of household consumption – as shown in Panel A of Figure 6 – although these measurements are biased – as shown in

[14]This finding is sufficient to reject the assumption of Berkson-type errors: $Y^* = Y^r + \epsilon^r$.

column (4) of Table 3. Predictions from diaries are also subject to a large variance, which is the takeaway message from Panel A of Figure 5. It follows that when aggregate statistics such as poverty rates or inequality are of interest, recall data yield more reliable figures than diary interviews.

On the bright side, diary errors present important similarities with classical errors in measurement. This conclusion can be reached by noticing that the error densities in Panel B of Figure 5 are reasonably stable across households with different levels of consumption. Column (5) of Table 3 also shows that MSE varies little across consumption percentiles. By taking logs, we conclude that $\log Y^d$ is centered below $\log Y^*$ and has variance approximately independent of $\log Y^*$, which is a finding with important implications for applied research.[15]

5.2 Consumption Measurements and the Frequency of Purchases

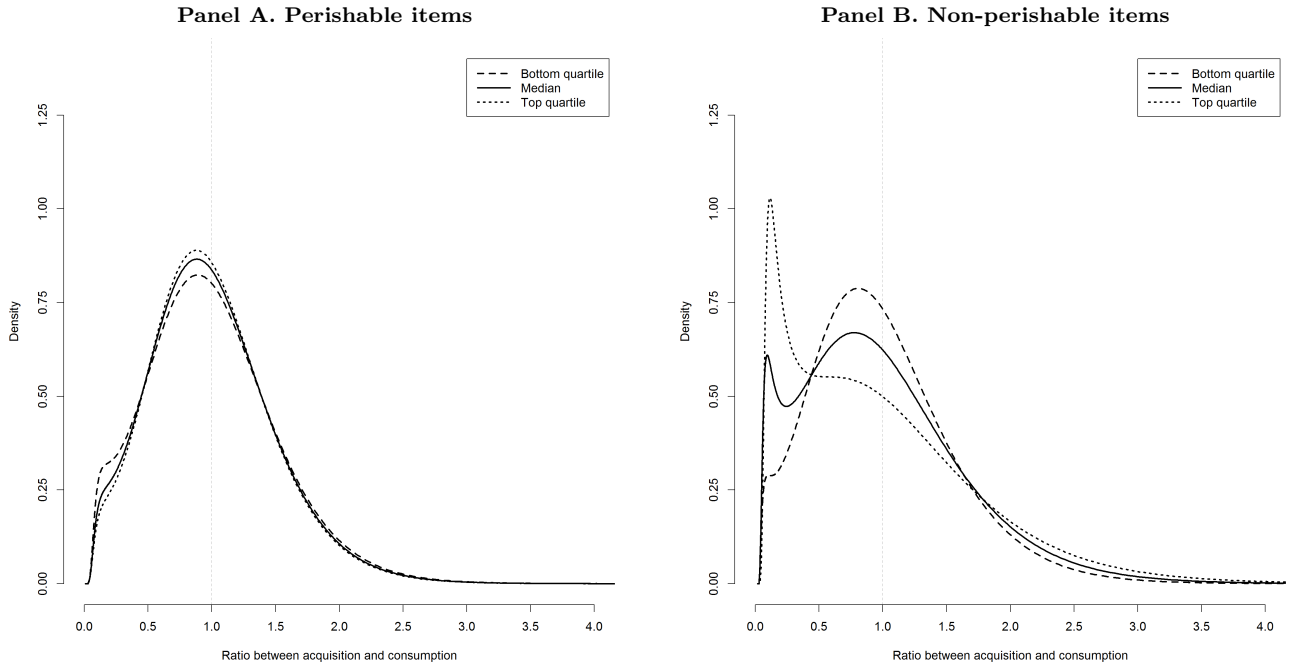
Large deviations from the mean in diary error distributions in Figure 5 can hardly be explained by the misreporting of diary entries, as respondents are assisted by frequent visits from enumerators. The most likely explanation is differences between acquisition in the diary week and the household's usual consumption Y^*. As mentioned in the Introduction, one possible reason for these differences is acquisition patterns arising from stock inflows or consumption out of existing stocks. The practical question then arises of whether diary errors would be substantially lower if one could adjust for stock inflows and outflows, as in Beegle et al. (2012) or Sharp et al. (2019). Although theory implies a negative answer unless household consumption is smoothed exactly across weeks (that is, the case $\tau = 1$ in the consumer demand model), this ultimately remains an empirical matter.

As we cannot observe stock inflows and outflows during the diary week, we replicate the analysis in the last section considering only the more perishable items for which consumption is arguably close to acquisition. Panel A of Figure 7 shows that diary acquisitions of perishable components of food consumption are characterized by large errors. Specifically, items entering our definition of consumption in Panel A are eggs, lamb, beef, chicken and fish, and the budget share devoted to these items is approximately 50% of total acquisition. Panel B of Figure 7 shows the results for the remaining items, which we label non-perishable for brevity. Diary errors Y^d/Y^* in Panel A present marginally lower dispersion at the top end of the consumption spectrum, as was the case in Figure 5 (note that Y^* in this panel now refers to usual consumption of perishable items). Errors can be large and lead to important under- or overstatements of the underlying consumption. For example, using a diary, the probability of attributing a value of consumption that is half ($Y^d/Y^* < 0.5$) or more than double ($Y^d/Y^* > 2$) the real value is 15% and 3.2%, respectively, for a household with median Y^* (the continuous line in Panel A).

The important implications of our approach for empirical work can be understood from Panel B of Figure 7, which shows how consumption from stocks or large acquisitions in the diary week affect the values of Y^d/Y^*. Distributions in this panel are centered at one – because of Assumption 1 – and strikingly bimodal, with spikes at zero signaling no acquisition in the week identified in the interview. For this group of items, the probability of attributing a value of consumption that is half ($Y^d/Y^* < 0.5$) or more than

[15] Specifically, our findings imply that $\log Y^* = \log Y^r + \log \epsilon^r$, where $\log \epsilon^r$ has constant variance and average below one, is a good specification to consider for empirical work.

Figure 7: Survey errors in acquisition diaries and frequency of purchases



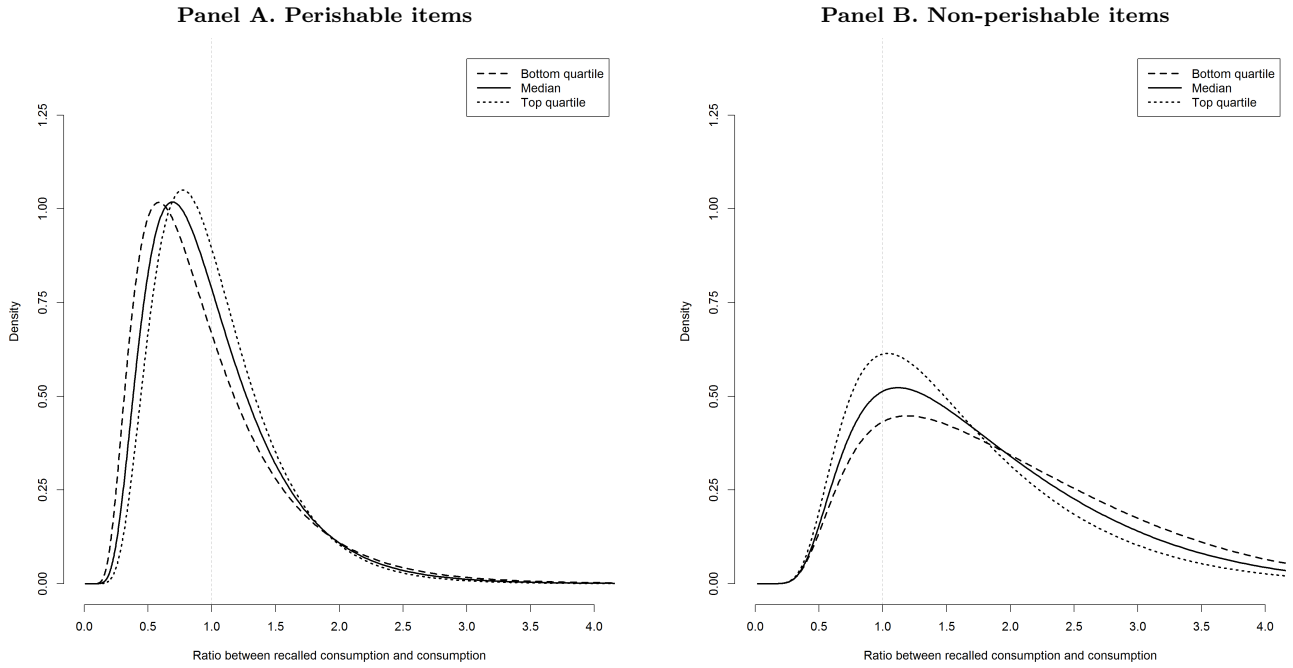
Note. This figure shows error distributions for a hypothetical household at three selected percentiles of usual consumption of perishable (Panel A) and non-perishable (Panel B) goods. The quantity Y^d/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from diaries, Y^d, and her usual consumption. No diary error implies a value of one for this ratio. Panel A considers the following goods: eggs, lamb, beef, chicken and fish (about 50% of the budget share). Panel B considers all remaining goods.

double ($Y^d/Y^* > 2$) the real value is 30.5% and 7.1%, respectively, for a household with median Y^* (the continuous line). Households in the top quartile of the consumption distribution of non-perishable items are those most likely to present no entries in their diary, as for these households the spike at zero is the highest. This finding is consistent with the fact that richer households might be able to stock up more and purchase items more infrequently.

A juxtaposition of densities in the two panels of Figure 7 sheds light on the anatomy of differences between acquisition and consumption in diary surveys. Assuming away misreporting of diary entries because of enumerators' visits, patterns in Panel B combine the effect of infrequent consumption and the effect of stock inflows and outflows during the diary week. The latter effect is eliminated in Panel A, where distributions become unimodal and with a lower probability of large positive deviations from the average. The comparison between panels should be taken with a grain of salt, since households need not maintain the same rank in the consumption distribution of perishable and non-perishable items. However, errors in Panel A demonstrate how perfect measurements of household consumption in a randomly selected week will still give misleading conclusions about actual well-being, which is a longer-run average of weekly intakes.

Recalled consumption of perishable items yields measurements of usual consumption that are more accurate than those obtained from a diary. In Panel A of Figure 8, the distributions of Y^r/Y^* suggest that recall errors are not too far from being zero on average. For example, the probability of underreporting ($Y^r/Y^* < 0.5$) or overreporting ($Y^r/Y^* > 2$) consumption is 12.2% and 4.6%, respectively, for a household

Figure 8: Survey errors in recalled consumption and frequency of purchases



Note. This figure shows error distributions for a hypothetical household at three selected percentiles of usual consumption of perishable (Panel A) and non-perishable (Panel B) goods. The quantity Y^r/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from the recall module, Y^r, and her usual consumption. No recall error implies a value of one for this ratio. Panel A considers the following goods: eggs, lamb, beef, chicken and fish (about 50% of the budget share). Panel B considers all remaining goods.

with a median level of consumption. A comparison with Panel B of the same figure reveals that the ability to recall consumption deteriorates for non-perishable items. Once again, the simplest story seems most likely: the respondent must backcast the monetary value of consumption, and this proves to be a relatively more difficult task for items that are acquired less frequently and bought in bulk. This interpretation is also supported by the fact that lower errors are observed for households at the top end of the consumption distribution.

6 Implications for the Design of Household Surveys

Is there an optimal assignment of households to diary and recall interviews? The error distributions documented above show that neither of these collection modes yields data that are of uniformly better quality. Consider a household with usual consumption equal to $Y^* = y^*$. If assigned to a diary, her food acquisition in the interview week would be a draw from $F_{Y^d|Y^*}[y|y^*]$. Similarly, a recall interview administered to the same household would yield a draw from $F_{Y^r|Y^*}[y|y^*]$. We ask which assignment rule across households would ensure the minimum loss of accuracy in raw data.[16]

Suppose that, at each value y^*, households are assigned a diary with probability $p(y^*)$. The data

[16]Others have considered the optimal design of surveys in terms of the trade-off between cost and survey errors (Manski, 2015, and Dominitz and Manski, 2017) or the optimal allocation of units to alternative treatments (Kitagawa and Tetenov, 2018, and Kasy and Sautmann, 2019).

distribution arising from this design is:

$$F_Y[y] = \int \left[F_{Y^d|Y^*}[y|y^*]\, p(y^*) + F_{Y^r|Y^*}[y|y^*]\,(1 - p(y^*)) \right] dF_{Y^*}[y^*], \tag{6}$$

the percent of survey participants filling out a diary being:

$$p \equiv \int p(y^*)\, dF_{Y^*}[y^*]. \tag{7}$$

For example, in a recall survey $p(y^*) = 0$ at all values y^*. We are interested in the effects of the collection mode on functionals of the distribution of usual consumption $F_{Y^*}[y^*]$. Possible choices for these functionals are quantiles, share below the poverty line, and the Gini coefficient. The difference:

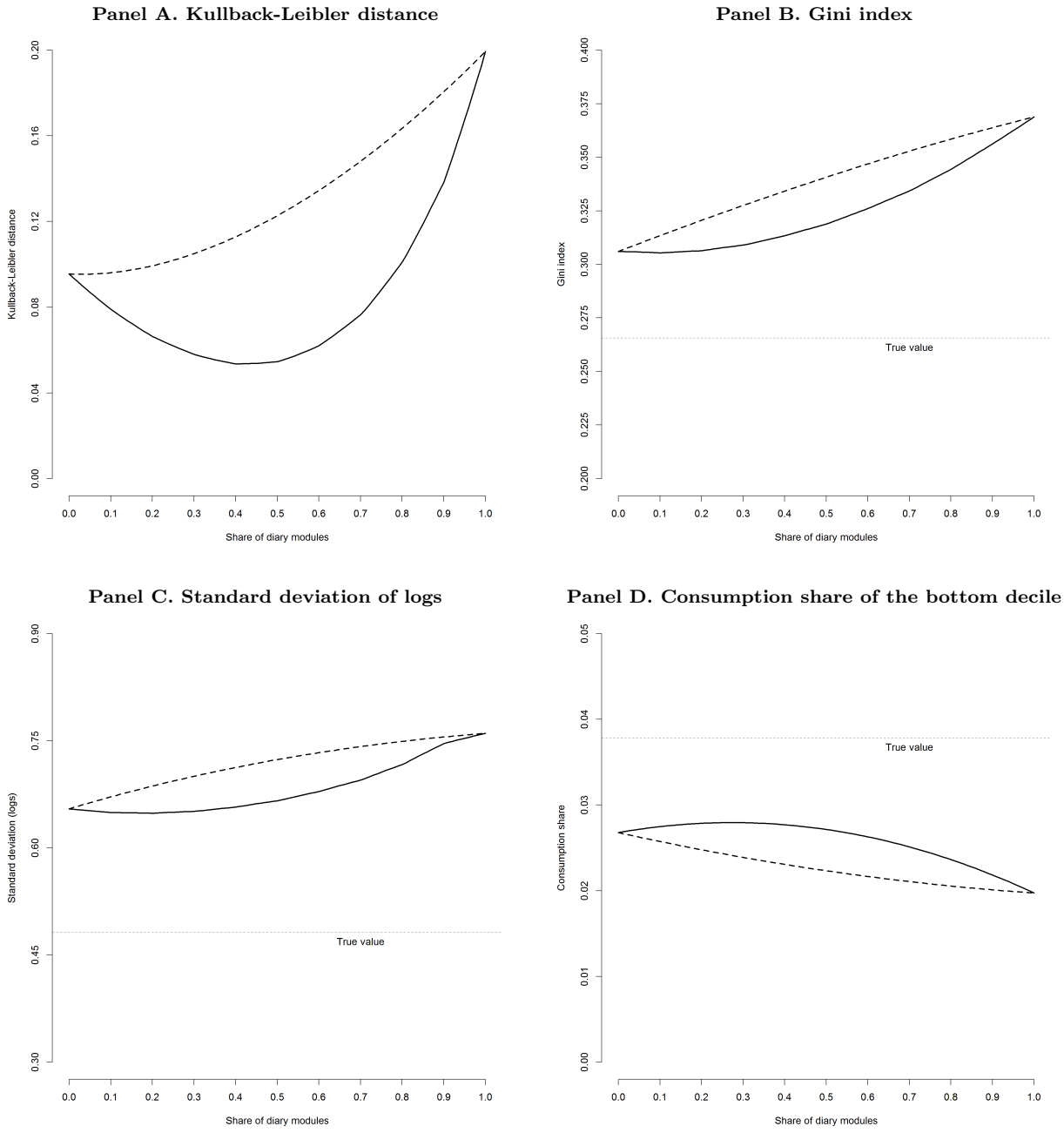
$$\nu(F_Y[y]) - \nu(F_{Y^*}[y^*]), \tag{8}$$

represents the distance between the true statistic, $\nu(F_{Y^*}[y^*])$, and the same statistic that would be computed under the assignment design described above, $\nu(F_Y[y])$. In this section, we find the configuration of weights $p(y^*)$ that minimizes the distance in (8) at any given choice for the value of p (the algorithm used to obtain the solution is described in the Appendix). In an effort to contain the computational burden, we use $F_{Y^d|Y^*}[y|y^*]$, $F_{Y^r|Y^*}[y|y^*]$ and $F_{Y^*}[y^*]$ as if they were *known* quantities and discard the additional variability arising from estimation errors.

Distributions retrieved from a recall survey are closer to the underlying distribution of usual consumption than in a diary survey. This can be seen in Panel A of Figure 9, where the statistic ν represents the Kullback-Leibler distance (KLD) from the consumption distribution and $\nu(F_{Y^*}[y^*]) = 0$. The continuous line here shows the value of (8) at different values of p. The value corresponding to $p = 0$ is the KLD from a survey with only recall questions. The panel conveys two important messages. First, as more resources are allocated to diaries, the KLD improves until it reaches the value $p = 0.42$. In particular, Appendix Figure A.1 shows that weights $p(y^*)$ at the value $p = 0.42$ are such that households with large values of Y^* should always be assigned to diaries. The second important message is that the KLD is larger at $p = 1$ than at $p = 0$. This implies that, in our case study, the data distribution obtained from a fully fledged recall survey has better properties than one obtained from diaries alone.

Distributions obtained from diary data yield worse measurements of poverty and inequality. This can be seen in the remaining panels of Figure 9, where the continuous lines represent alternative definitions for the statistic $\nu(F_Y[y])$: the Gini coefficient (Panel B), the standard deviation of logged consumption (Panel C), and the share of consumption for households in the bottom decile (Panel D). The horizontal lines in panels represent the value of the true statistic $\nu(F_{Y^*}[y^*])$. Diaries yield more disperse consumption distributions, as shown in Panel B and Panel C. Moreover, the Gini coefficient and the standard deviation computed from a recall survey are closer to their true values than one would obtain by using only diaries. It is also clear that allocating more resources to diaries would improve inequality measurement only until approximately the value $p = 0.10$. This suggests that a fully fledged recall survey is almost as close to the optimal design when inequality measurement is of concern. Panel D yields very similar conclusions

Figure 9: Optimal allocation to modes of data collection



Note. The horizontal axis shows the share p of diary interviews. The value $p = 0$ corresponds to a recall-only survey (no diaries). The value $p = 1$ corresponds to a fully-fledged diary survey. The continuous lines are obtained by allocating to diaries a share p of households using their usual consumption Y^* - see equations (6) and (7). The dashed lines are obtained by allocating to diaries a random share p of households independently of their usual consumption. Panel A shows the Kullback-Leibler distance between the distribution of usual consumption and the empirical distributions under alternative allocations indexed to p. The statistics in the remaining panels use the distribution of usual consumption and the empirical distributions under alternative allocations indexed to p: Gini index (Panel B), standard deviation of logs (Panel C), consumption share of the bottom decile (Panel D).

for poverty measurement.

We conclude our analysis by showing that a recall survey is the best option even in the worst case scenario when no information on Y^* is available.[17] This can be seen from the dashed lines in Figure 9, which are obtained by setting $p(y^*) = p$ in (6) and allocating to diaries a *random* share p of households independently of their consumption. The lines in the four panels suggest that there is no gain from mixing diary and recall interviews without a selective allocation of households that depends on their consumption.

The take-away message from this section is that we do not find evidence of any loss in accuracy from using recall questions compared to a diary. Moreover, the above definition of optimal design does not take into account costs and budget constraints in the optimization. Our conclusions on the value of recall surveys would be strengthened given the much higher costs of running a diary survey.

7 Conclusion

Using a large-scale randomization in Iraq, we found little empirical support for the idea that diaries yield better quality data for assessing household welfare. Diaries provide a more reliable measurement of usual consumption *averages*, and we have shown that the cognitive errors arising from the process of recalling consumption lead to overstated actual consumption. However, when inequality and poverty measurements are of interest, the benefits of diaries are far less clear-cut. Diary measurements, despite being correct on average, have large variance. We have demonstrated that this is not the consequence of measurement errors but mostly a reflection of heterogeneous frequency of consumption across households. We have found that recall modules provide a more reliable mode of data collection for inequality and poverty. The use of surveys with both diary and recall interviews can yield improved measurements because these two collection modes work best for eliciting consumption in different segments of the distribution. Nevertheless, our calculations have shown that a fully fledged recall survey can yield inequality measurements that are not too different from those that would be obtained from the optimal mix of diary and recall interviews.

The first implication of our research is that the loss in accuracy from using recall questions is minimal compared to the higher costs of using diaries. This finding provides an empirical justification for considering a transition to recall modules in household surveys in developing countries. However, more research is needed to assess what makes a good recall module, given that the length of the recall list will affect the propensity of respondents to engage in the survey. The answer to this question could be given, for example, by randomizing households to recall modules of different length and using our research design. The decision between using diaries or recall modules is not confined to developing economies. For example, the problems associated with the Consumer Expenditure Surveys in the United States have intensified the call for a redesign of the survey, which is underway following the recommendations of a specially appointed panel (the Gemini project).

Diary surveys should collect information on the purchasing behavior of households to correct for the unwelcome effects of infrequent purchases on inequality measurement – which is the second implication

[17] Allocation based on proxies of Y^*, such as income or residence in areas with homogeneous compositions of household demographics, falls between this case and the one above.

of this research. Respondents should be asked to estimate the number of purchases made over a fixed reference period, anchoring questions on consumption rather than on acquisitions (as in the 2016 Barbados Survey of Living Conditions and the 2017/18 Jordan Household Income and Expenditure Survey). If information on purchasing behavior is available, simple models can be used to estimate the effects of using diaries on inequality measurements.

A third implication of our findings is that surveys should be designed to elicit repeated measurements from the same respondents. In our case, for example, a recall module was administered to all households before starting the diary week. The same setting can be found in other important family expenditure surveys, such as in Canada and the United States. The methodology we have presented can be applied to any of these contexts to study how consumption inequality has evolved over time, raising the problem of how our approach can be extended to allow for longitudinal information. We hope to address this and some of the related problems in future research.

References

Aguiar, Mark and Mark Bils, "Has Consumption Inequality Mirrored Income Inequality?," *American Economic Review*, September 2015, *105* (9), 2725–56.

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang, "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," Working Paper 26463, National Bureau of Economic Research November 2019.

Attanasio, Orazio P. and Luigi Pistaferri, "Consumption Inequality," *Journal of Economic Perspectives*, May 2016, *30* (2), 3–28.

Backiny-Yetna, Prospère, Diane Steele, and Ismael Yacoubou Djima, "The impact of household food consumption data collection methods on poverty and inequality measures in Niger," *Food Policy*, 2017, *72*, 7 – 19.

Barron, Andrew R. and Chyong-Hwa Sheu, "Approximation of Density Functions by Sequences of Exponential Families," *The Annals of Statistics*, 1991, *19* (3), 1347–1369.

Battistin, Erich, "Errors in survey reports of consumption expenditures," IFS Working Paper W03/07, Institute for Fiscal Studies April 2003.

_ **and Mario Padula**, "Survey instruments and the reports of consumption expenditures: evidence from the consumer expenditure surveys," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2016, *179* (2), 559–581.

_ **, Richard Blundell, and Arthur Lewbel**, "Why Is Consumption More Log Normal than Income? Gibrat's Law Revisited," *Journal of Political Economy*, 2009, *117* (6), 1140–1154.

Bee, Adam, Bruce D. Meyer, and James X. Sullivan, "The Validity of Consumption Data: Are the Consumer Expenditure Interview and Diary Surveys Informative?," in "Improving the Measurement of Consumer Expenditures," University of Chicago Press, February 2013, pp. 204–240.

Beegle, Kathleen, Joachim De Weerdt, Jed Friedman, and John Gibson, "Methods of household consumption measurement through surveys: Experimental results from Tanzania," *Journal of Development Economics*, 2012, *98* (1), 3 – 18.

Browning, Martin and Thomas Crossley, "Are Two Cheap, Noisy Measures Better Than One Expensive, Accurate One?," *American Economic Review*, May 2009, *99* (2), 99–103.

Brzozowski, Matthew, Thomas F. Crossley, and Joachim K. Winter, "A comparison of recall and diary food expenditure data," *Food Policy*, 2017, *72*, 53 – 61.

Carroll, Christopher D., Thomas F. Crossley, and John Sabelhaus, *Improving the Measurement of Consumer Expenditures*, University of Chicago Press, 2015.

Coibion, Olivier, Yuriy Gorodnichenko, and Dmitri Koustas, "Consumption Inequality and the Frequency of Purchases," *American Economic Journal: Macroeconomics (forthcoming)*, 2020.

Comerford, David, Liam Delaney, and Colm Harmon, "Experimental Tests of Survey Responses to Expenditure Questions," *Fiscal Studies*, 2009, *30* (3/4), 419–433.

Crossley, Thomas F. and Joachim K. Winter, "Asking Households about Expenditures: What Have We Learned?," in "Improving the Measurement of Consumer Expenditures," University of Chicago Press, July 2014, pp. 23–50.

Deaton, Angus and Margaret Grosh, "Consumption," in Margaret Gosh and Paul Glewwe, eds., *Designing Household Survey Questionnaires for Developing Countries*, The World Bank, 2000, chapter 5, pp. 91–134.

_ **and Salman Zaidi**, "Guidelines for Constructing Consumption Aggregates for Welfare Analysis," Technical Report, World Bank, Washington, DC 2002.

Dominitz, Jeff and Charles F. Manski, "More Data or Better Data? A Statistical Decision Problem," *The Review of Economic Studies*, 2017, *84* (4), 1583–1605.

FAO and The World Bank, "Food data collection in Household Consumption and Expenditure Surveys. Guidelines for low- and midele-income countries.," Technical Report, World Bank 2018. Rome.

Gieseman, Raymond, "The Consumer Expenditure Survey: Quality control by comparative analysis," *Monthly Labor Review*, 1987, *110* (3), 8 – 14.

Grenander, Ulf, *Abstract Inference*, Wiley, 1981.

Hoderlein, Stefan and Joachim Winter, "Structural measurement errors in nonseparable models," *Journal of Econometrics*, 2010, *157* (2), 432 – 440.

Hu, Yingyao and Susanne M. Schennach, "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 2008, *76* (1), 195–216.

Hyslop, R. and Guido W. Imbens, "Bias from Classical and Other Forms of Measurement Error," *Journal of Business & Economic Statistics*, 2001, *19* (4), 475–481.

Kasy, Maximilian and Anja Sautmann, "Adaptive Treatment Assignment in Experiments for Policy Choice," *working paper,*, 2019.

Kay, J.A., M.J. J. Keen, and C.N. N. Morris, "Estimating Consumption from Expenditure data," *Journal of Public Economics*, 1984, *23* (1-2), 169–181.

Kitagawa, Toru and Aleksey Tetenov, "Who should be treated? Empirical welfare maximization methods for treatment choice," *Econometrica*, May 2018, *86* (2), 591–616.

Lyberg, Lars and Daniel Kasprzyk, "Data Collection Methods and Measurement Error: An Overview," in "Measurement Errors in Surveys," Wiley-Blackwell, 2004, chapter 13, pp. 235–257.

Manski, Charles F., "Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern," *Journal of Economic Literature*, September 2015, *53* (3), 631–53.

Meghir, Costas and Jean-Marc Robin, "Frequency of purchase and the estimation of demand systems," *Journal of Econometrics*, 1992, *53* (1), 53 – 85.

Neter, John and Joseph Waksberg, "A Study of Response Errors in Expenditures Data from Household Interviews," *Journal of the American Statistical Association*, 1964, *59* (305), 18–55.

Newey, Whitney K. and James L. Powell, "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, sep 2003, *71* (5), 1565–1578.

Peterson Zwane, Alix, Jonathan Zinma et al., "Being surveyed can change later behavior and related parameter estimates," *Proceedings of the National Academy of Sciences*, 2011, *108* (5), 1821–1826.

Pudney, Stephen, "Heaping and leaping: survey response behaviour and the dynamics of self-reported consumption expenditure," Technical Report, ISER 2008.

Schennach, Susanne M., "Measurement Error in Nonlinear Models - A Review," in Daron Acemoglu, Manuel Arellano, and Eddie Dekel, eds., *Advances in Economics and Econometrics: Tenth World Congress*, Vol. 3 of *Econometric Society Monographs*, Cambridge University Press, 2013, pp. 296–337.

Sharp, Michael, Bertrand Buffiere, Kristen Himelein, and John Gibson, "Effects of Data Collection Methods on Estimated Household Consumption and Poverty, and on Survey Costs: Evidence from an Experiment in the Marshall Islands," in "Working Paper" Paper prepared for the IARIW-World Bank Conference 2019.

Silberstein, Adriana R. and Stuart Scott, "Expenditure Diary Surveys and Their Associated Errors," in "Measurement Errors in Surveys," Wiley-Blackwell, 2004, chapter 16, pp. 303–326.

Wilhelm, Daniel, "Testing for the Presence of Measurement Error," Working Paper CWP48/19, CEMMAP 2019.

Online Appendix

Acquisition Variance with One Item

Let acquisition for one good be defined as in equation 1. The variance of Y^d is:

$$\begin{aligned} Var(Y^d) &= Var\left(Y^* \frac{N}{N^*}\right), \\ &= E\left[\left(\frac{Y^*}{N^*}\right)^2 Var(N|N^* = n^*, Y^* = y^*)\right] + Var(Y^*). \end{aligned} \tag{A.1}$$

To fix ideas, consider the case of:

- independent and identically distributed time intervals between acquisitions at a rate of N^* per week, which implies $N|N^*, Y^* \sim Poisson(N^*)$, and
- no heterogeneity in purchasing frequency across households.

Under these assumptions we can rewrite the last equation as:

$$\begin{aligned} Var(Y^d) &= Var(Y^*) + E\left[\frac{Y^{*2}}{N^*}\right], \\ &= Var(Y^*) + \frac{Var(Y^*)}{N^*} + \frac{E[Y^*]^2}{N^*}, \end{aligned}$$

which implies:

$$\frac{Var(Y^d)}{Var(Y^*)} = 1 + \frac{1}{N^*} + \frac{1}{CV(Y^*)^2 N^*}, \tag{A.2}$$

where $CV(Y^*)$ is the coefficient of variation of consumption.

Frequency of Purchases with Multiple Items

Consider for simplicity the case of two items i and j, as the generalization to multiple items follows straightforwardly. A frequency of purchase model for a vector of goods implies (see Meghir and Robin, 1992):

$$E[N_i|N_i^* = n_i^*, Y_i^* = y_i^*, N_j^* = n_j^*, Y_j^* = y_j^*] = n_i^*, \tag{A.3}$$
$$E[N_j|N_i^* = n_i^*, Y_i^* = y_i^*, N_j^* = n_j^*, Y_j^* = y_j^*] = n_j^*. \tag{A.4}$$

We do not entertain the possibility of corner solutions, as food consumption is a necessity and the rate of consumption Y^* must be positive. Starting from equation (3), it is immediate to see that the case of perfectly dependent purchases across items ($N_i^* = N_j^*$ and $N_i = N_j$) implies (2). The same conclusion holds in the more general case. To see this, we write equation (3) in the case of two items:

$$Y^d = Y^* \left(b_i^* \frac{N_i}{N_i^*} + b_j^* \frac{N_j}{N_j^*}\right),$$

and obtain:

$$E[Y^d|N_i^* = n_i^*, Y_i^* = y_i^*, N_j^* = n_j^*, Y_j^* = y_j^*] = Y^* \left(b_i^* + b_j^*\right) = Y^*,$$

which implies (2) after integration. The extension to the case of multiple items follows by analogy.

Acquisition Variance with Multiple Items

Let acquisition for I goods be defined as in equation (3) by:

$$Y^d = \sum_{i=1}^{I} Y_i^* \frac{N_i}{N_i^*},$$

where Y_i^* is consumption for item i, N_i^* is the average number of purchases in one week for item i and N_i is the observed number of purchases in the diary week. The variance of Y^d can be written as:

$$\begin{aligned} Var(Y^d) &= Var\left(\sum_{i=1}^{I} Y_i^* \frac{N_i}{N_i^*}\right), \\ &= \sum_{i=1}^{I}\sum_{j=1}^{I} Cov\left(Y_i^* \frac{N_i}{N_i^*}, Y_j^* \frac{N_j}{N_j^*}\right). \end{aligned} \tag{A.5}$$

Note that

$$\begin{aligned} Cov\left(Y_i^* \frac{N_i}{N_i^*}, Y_j^* \frac{N_j}{N_j^*}\right) &= E\left[Y_i^* \frac{N_i}{N_i^*} Y_j^* \frac{N_j}{N_j^*}\right] - E\left[Y_i^* \frac{N_i}{N_i^*}\right] E\left[Y_j^* \frac{N_j}{N_j^*}\right] \\ &= E\left[\frac{Y_i^* Y_j^*}{N_i^* N_j^*} \times \right. \\ &\quad \left. E\left[N_i N_j | N_i^* = n_i^*, N_j^* = n_j^*, Y_i^* = y_i^*, Y_j^* = y_j^*\right]\right] \\ &- E[Y_i^*]E[Y_j^*] \\ &= E\left[\frac{Y_i^* Y_j^*}{N_i^* N_j^*}\sigma_{ij}\right] + E[Y_i^* Y_j^*] - E[Y_i^*]E[Y_j^*] \\ &= Cov(Y_i^*, Y_j^*) + E\left[\frac{Y_i^* Y_j^*}{N_i^* N_j^*}\sigma_{ij}\right], \end{aligned} \tag{A.6}$$

where $\sigma_{ij} = Cov(N_i, N_j | N_i^*, N_j^*, Y_i^*, Y_j^*)$ and the result follows because of (A.3) for $i = 1, \ldots, I$. Substituting (A.6) into (A.5) yields:

$$\begin{aligned} Var(Y^d) &= \sum_{i=1}^{I}\sum_{j=1}^{I}\left\{Cov(Y_i^*, Y_j^*) + E\left[\frac{Y_i^* Y_j^*}{N_i^* N_j^*}\sigma_{ij}\right]\right\}, \\ &= Var(Y^*) + \sum_{i=1}^{I}\sum_{j=1}^{I} E\left[\frac{Y_i^* Y_j^*}{N_i^* N_j^*}\sigma_{ij}\right], \end{aligned} \tag{A.7}$$

which provides the general form of $Var(Y^d)$ when aggregation is over I goods.

To fix ideas, consider the case of:

- independent and identically distributed time intervals between acquisitions at a rate of N_i^* per week, for good i, which implies $N_i | N_i^*, Y_i^* \sim Poisson(N_i^*)$ for $i = 1, \ldots, I$, and
- no heterogeneity in purchasing frequency across households, and
- constant correlation ρ between N_i and N_j.

These assumptions allow to simplify the expression above as follows:

$$\begin{aligned} Var(Y^d) &= Var(Y^*) + \sum_{i=1}^{I} E\left[\left(\frac{Y_i^*}{N_i^*}\right)^2 \sigma_{ii}\right] \\ &+ \rho \sum_{i=1}^{I}\sum_{j\neq i} E\left[\frac{Y_i^* Y_j^*}{N_i^* N_j^*}\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}\right], \\ &= Var(Y^*) + \sum_{i=1}^{I} E\left[\frac{Y_i^{*2}}{N_i^*}\right] + \rho \sum_{i=1}^{I}\sum_{j\neq i} E\left[\frac{Y_i^* Y_j^*}{\sqrt{N_i^*}\sqrt{N_j^*}}\right], \\ &= Var(Y^*) + \sum_{i=1}^{I} \frac{1}{N_i^*} E\left[Y_i^{*2}\right] + \rho \sum_{i=1}^{I}\sum_{j\neq i} \frac{1}{\sqrt{N_i^*}\sqrt{N_j^*}} E\left[Y_i^* Y_j^*\right]. \end{aligned}$$

When acquisitions for different goods are independent we have $\rho = 0$ for $i \neq j$, which implies:

$$\begin{aligned} Var(Y^d) &= Var(Y^*) + \sum_{i=1}^{I} \frac{1}{N_i^*} E\left[Y^2\right], \\ &= Var(Y^*) + \sum_{i=1}^{I} \frac{1}{N_i^*} Var(Y_i^*) + \sum_{i=1}^{I} \frac{1}{N_i^*} E[Y_i^*]^2, \end{aligned}$$

By rearranging terms in the last expression we get:

$$\frac{Var(Y^d)}{Var(Y^*)} = 1 + \sum_{i=1}^{I} \frac{1}{N_i^*} \frac{Var(Y_i^*)}{Var(Y^*)} + \sum_{i=1}^{I} \left(\frac{E[Y_i]}{E[Y^*]}\right)^2 \frac{1}{CV(Y^*)^2 N_i^*}. \tag{A.8}$$

Manipulation of Purchases in the Diary Week

If purchases in the diary week are manipulated, we have that:

$$E[N_i | N_i^*, Y_1^*, \ldots, Y_I^*] = \alpha N_i^*,$$

where α is a random variable across households (the case of item-specific α's leads to conditions similar to the one below). Using the expression above one can write:

$$E[Y^d | N_i^* = n_i^*, Y_i^* = y_i^*, N_j^* = n_j^*, Y_j^* = y_j^*] = Y^* E[\alpha | N_i^* = n_i^*, Y_i^* = y_i^*, N_j^* = n_j^*, Y_j^* = y_j^*],$$

which also implies:

$$E[Y^d | Y^*] = Y^* E\left[\alpha | Y^*\right].$$

The variable α is household specific and accomodates for the possible effects on the household's purchasing behaviour as a result of being interviewed. The condition $E\left[\alpha | Y^*\right] = 1$ ensures that diary errors are zero on average.

Optimal allocation of diary and recall interviews

We describe here how weights $p(y^*)$ in Section 6 were determined to minimize the Kullback-Leibler distance (KLD) between the observed distribution and the consumption distribution estimated in Figure 1. We started by writing $p(y^*)$ as:

$$p(y^*; \boldsymbol{\alpha}) = \sum_{q=1}^{10} \alpha_q \mathbf{1}(y_{(q-1)}^* \leq y^* < y_{(q)}^*),$$

where $y^*_{(0)}, y^*_{(1)}, \ldots, y^*_{(10)}$ are the deciles of the distribution of Y^* and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{10})'$. The last expression defines a piecewise-constant function for $p(y^*)$, with α_i representing the probability of being assigned a diary module for individuals in the i-th decile of the distribution of consumption. The optimal assignment is obtained by solving the following problem:

$$\begin{aligned} \hat{\boldsymbol{\alpha}} &= \underset{\boldsymbol{\alpha}}{\text{argmax}}\ \nu(F_Y(y;\boldsymbol{\alpha})), \\ &\text{s.t.}\quad 0 \leq \alpha_i \leq 1, \text{ for } i = 1, \ldots, 10, \end{aligned}$$

where $\nu(F_Y(y))$ is the KLD of the distribution of Y from the consumption distribution Y^*:

$$\nu(F_Y(y)) = \int_0^\infty f_Y(y) \log\left(\frac{f_Y(y)}{f_{Y^*}(y)}\right) dy,$$

and:

$$F_Y(y) = \int \left[F_{Y^d|Y^*}(y^d|y^*)p(y^*;\boldsymbol{\alpha}) + F_{Y^r|Y^*}(y^r|y^*)(1 - p(y^*;\boldsymbol{\alpha}))\right] dF_{Y^*}(y^*).$$

This defines a constrained optimization problem over $\boldsymbol{\alpha}$. The outcome of such optimization is shown in Figure A.1. Panel B reports the estimated $p(y^*;\hat{\boldsymbol{\alpha}})$, which implies that the optimal survey design assigns diary modules with probability one to households above the sixth decile of the consumption distribution and recall modules with probability one to the remaining households. We therefore have:

$$\int_0^\infty p(y^*;\hat{\boldsymbol{\alpha}}) dF_{Y^*}(y^*) = 0.42.$$

The improvement in terms of observed consumption distribution can be seen from Panel A, where the consumption density is plotted alongside the observed densities obtained by assigning only diaries, only recall modules or the optimal survey design.

We also considered the solution to the following problem:

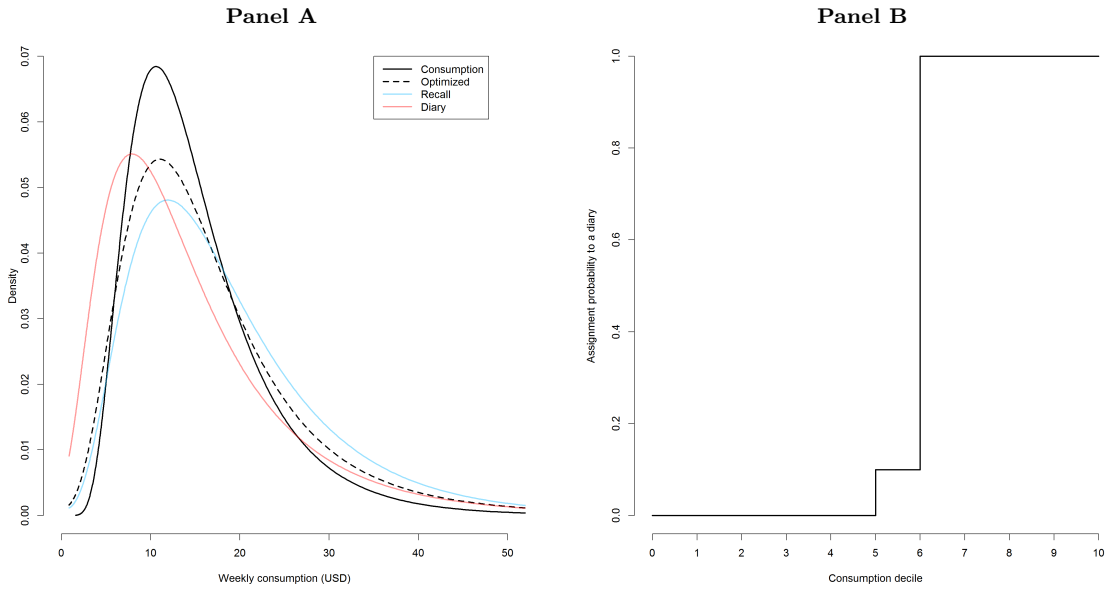
$$\begin{aligned} \hat{\boldsymbol{\alpha}}_p &= \underset{\boldsymbol{\alpha}}{\text{argmax}}\ \nu(F_Y(y;\boldsymbol{\alpha})) \\ &\text{s.t.}\quad 0 \leq \alpha_i \leq 1, \text{ for } i = 1, \ldots, 10 \\ &\text{s.t.}\quad \int_0^\infty p(y^*;\boldsymbol{\alpha}) dF_{Y^*}(y^*) = p, \end{aligned}$$

which yields the optimal allocation of a share p of households to diaries. Functionals of the observed distributions under alternative choices of p, i.e.,

$$F_Y(y;p) = \int \left[F_{Y^d|Y^*}(y^d|y^*)p(y^*;\hat{\boldsymbol{\alpha}}_p) + F_{Y^r|Y^*}(y^r|y^*)(1 - p(y^*;\hat{\boldsymbol{\alpha}}_p))\right] dF_{Y^*}(y^*)$$

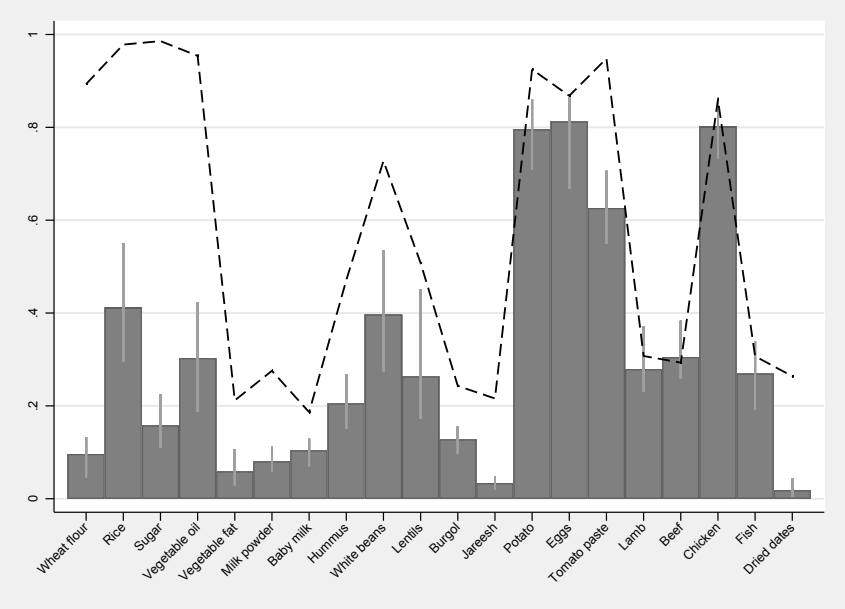
are reported in Figure 9.

Figure A.1: Optimal survey design



Note. Panel A of this figure presents densities for the treatment sample which were fitted using: (a) raw diary data (Diary); (b) raw recall data (Recall); (c) the model in Section 4 to obtain the usual consumption distribution (Consumption); (d) the allocation in Section 6 at the minimum value of the Kullback-Leibler distance (Optimized). The treatment sample consists of households randomized to the recall-module experiment. Panel B shows the weights $p(y^*)$ in equation (6) corresponding to the minimum value of the Kullback-Leibler distance. Acquisitions and valuations of recalled consumption, in Iraqi dinars, are reported in US dollars (USD) on the horizontal axis.

Figure A.2: Recalled consumption and diary acquisition across food items



Note. This figure compares the share of households with positive spending in the diary week (in bars) to the share of households self-reporting positive consumption in the recall module (the dashed line). The former quantity is computed using all households. The vertical line around bars shows how the share with positive spending varies across survey months. The 20 consumption groups used in the recall module are reported on the horizontal axis.