2019

IARIW-World Bank

Special IARIW-World Bank Conference "New Approaches to Defining and Measuring Poverty in a Growing World" Washington, DC, November 7-8, 2019

The Insights and Illusions of Consumption Measurement: Evidence from a Large Scale Randomization

Erich Battistin

Michele De Nadai

Nandini Krishnan

Paper Prepared for the IARIW-World Bank Conference

Washington, DC, November 7-8, 2019

Session 1A: Methods for Measuring Consumption and Poverty

Time: 11:00 – 13:00, September 7

The Insights and Illusions of Consumption Measurement: Evidence from a Large Scale Randomization[*]

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

Michele De Nadai[‡]
University of New South Wales

Nandini Krishnan[§]
World Bank

June 2019

Abstract

We challenge the belief that acquisition diaries yield the most accurate measurement of household consumption. Using a large-scale randomization in Iraq that assigns households to survey modules with both recall and diary measurements, our novel strategy non-parametrically unveils the distribution of errors in both measurements and the distribution of true underlying consumption. Identification stems from differences between experimentally-formed groups of respondents to diary and recall interviews, combined with diary-recall differences in measurements from the same households. Our assumptions hinge on the same survey design employed in many national household surveys, like in Canada and the United States, and do not need diaries to be error-free or have classical errors. We find little empirical support for concluding that diaries outperform recall measurements, and offer new insights to interpret and reconcile differences between consumption measurements in the cross-section and over time. Diary errors are of greatest concern for studying the distributional aspects of welfare, as differences between *expenditure* and *consumption* rankings for poverty or inequality arise because of heterogeneous purchasing behavior across households. We devise practical suggestions for the design of consumption modules in household surveys that would yield the closest measurement to the true underlying consumption.

JEL Classification: C81, D31, D63, E21, I32
Keywords: Household Surveys, Measurement of Inequality and Poverty, Modes of Data Collection

[*] Preliminary draft. The manuscript benefited from comments by Beatriz Godoy, Juan Munoz, Sergio Olivieri and Tara Vishwanath. 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

The measurement of economic welfare is anchored in incomes or consumption expenditures. Food represents a large share of total expenditure in low-income and middle-income countries, which rely on household surveys to measure and monitor welfare. Global monitoring of extreme poverty relies on food-based estimates of welfare, such as the Cost of Basic Needs (CBN) approach. Information on food consumption is therefore central to developing indicators of poverty, food security, nutrition, health and for policy analysis. Means-tested interventions in most developing countries also rely on consumption-based measures of welfare, of which food is a dominant component. A number of programs explicitly target food consumption and nutrition, ranging from school meals, fortified foodstuffs, price controls, and food subsidies, including the Supplemental Nutrition Assistance Program (SNAP) in the United States. Moreover, food is the only expenditure category with a long panel component in household surveys such as the Panel Study of Income Dynamics (PSID) or the British Household Panel Study (BHPS). Because of its centrality to policy and welfare measurement, a large body of the literature has questioned the accuracy of alternative modes of collecting information on food consumption.

We revisit the issue of whether diaries are the most appropriate benchmark at a time when developing countries have increasingly moved to recall. Contrary to several influential studies discussed below, we find that the presumption that diaries yield better data is just an illusion. Diaries have become the workhorse for benchmarking comparisons of household measurements from alternative collection modes. Diaries are self-administered surveys in which respondents must report expenditure incurred over a short period, usually one or two weeks. Completion involves a reasonable standard of literacy and commitment. To ensure accurate measurement, mid-term or more frequent visits from interviewers has become the recommended norm. Beegle et al. (2012) randomize households in Tanzania to assisted individual 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). Accuracy has been a longstanding concern not only in developing countries. Brzozowski et al. (2017) compare recall questions on food spending from the Canadian Food Expenditure Survey to expenditure diaries. Battistin and Padula (2016) and others do so using the Consumer Expenditure Surveys (CEX) in the United States. In these instances too, the accuracy of survey reports is assessed against diaries. The idea that assisted diaries on food spending yield better data is widespread, as they arguably minimize various types of reporting errors, including recall errors, telescoping, rule of thumb errors, and leave out errors (Lyberg and Kasprzyk, 2011).

The assumption that diaries provide an error-free benchmark is done for convenience but has little statistical justification. Diaries are far from perfect. Most surveys are characterized by differences in reporting within the diary period, usually with a declining pattern over time (Silberstein and Scott, 2011). Possible explanations are 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 about persons other than the respondent yields figures different from those obtained from individual diaries (Beegle et al., 2012). We also know that the ratio between food totals from household surveys and national accounts can be considerably lower using diaries than recall questions (Gieseman, 1987, and Bee et al., 2013). Moreover, diaries are expensive: regular visits to ensure quality standards and data processing boosts costs.

Our view is that data quality can be a red herring in the study of the distributional aspects of welfare using diaries. One key aspect here is the difference between *consumption* and *spending*, which has been largely overlooked in empirical work. Household surveys elicit consumption or expenditure on food with the idea that they should yield similar figures. With the increasing shift towards recall, most surveys in developing countries focus on food consumed, the monetary value of consumption, and the sources from which food was acquired.[1] Recall modules ask households to report on food consumption during the specified period, and to value that consumption, irrespective of source, at market prices. Diaries instead measure household spending for market acquisitions, and market-valued consumption for gifts, self-production and other non-market sources. The comparison between diary and recall data may be misleading if the former source primarily measures spending and the latter consumption (as in Beegle et al., 2012, among others). The frequency of purchases may vary across food items, and differentially across socio-economic groups. Grain products such as rice, wheat and oil are often purchased in bulk and stored but may not be affordable for households relying on irregular and uncertain incomes. Lower frequency of shopping may be associated with larger expenditures at each trip. While assumptions can be made to claim that spending and consumption are the same on *average*, as we discuss below, heterogeneity in the frequency of purchases will typically increase dispersion and has an impact on indicators of poverty and inequality (in the cross-section and over time). Assuming that diaries measure expenditures without errors is not enough to account for these problems.

Differences between experimentally-formed groups of respondents to diary and recall interviews are, in general, only partially informative about survey errors. The assumption that spending and consumption have the same average, combined with the assumption of perfect recording of diaries, identify the *average* error committed by recall respondents. Unless spending coincides with consumption, however, differences in welfare indicators (e.g., the Gini coefficient) between the two experimental groups are not solely the result of data quality. This is true also when diary entries are complete, and entries have no errors. For example, even random deviations of diary spending from household consumption would raise the density of diary data where it is convex (e.g., in the tails) and flatten it where it is concave (e.g., at its mode), increasing dispersion (as implied by prior work in Chesher and Schluter, 2002). Knowing that the survey mode affects indicators of poverty and inequality in the cross-section (Beegle et al., 2012) and over time (Battistin, 2003, Attanasio and Pistaferri, 2016, and Coibion et al., 2017), empirical questions arise of how the true underlying household consumption can be computed, and if any mismeasurement from alternative survey modes can be quantified.

How do we address these empirical questions? We leverage a unique large-scale experiment in Iraq designed for the Household and Socio-Economic Survey (IHSES) in 2012. The survey ran continuously, year-long, and was administered to about 25,000 households across the country. All households filled out a 7 day diary on their spending patterns, assisted by enumerators during regular visits. One-third of households in each primary sampling unit were randomized to an additional survey module, administered before the diary, asking them to recall food consumption in the last 7 days. Both modules had the same respondent, the household head or the most informed adult. The size of the "treatment" sample, of roughly 8,000 households, is therefore of a different order of magnitude compared to other studies in developing countries. The household diary employed was standard, with one respondent recording on behalf of the household.[2] It is of the acquisition type used in different contexts, asking about quantities purchased from market sources and quantities consumed from non-market sources and their market valuations (Albania 2002 and 2005, Ghana 1991 and 1998, Iraq 2007, Serbia 2002, 2003 and 2007, for instance, and Canada, United States and United Kingdom). The recall module was specifically designed to inform the national statistics agency on the transition from diary to recall, which is planned for 2019-2020. A list of recall expenditure groups was selected based on an assessment of their importance in household food expenditure shares and on how commonly they were reported,

[1] The World Bank's Living Standards Measurement Study (LSMS) survey finder includes a partial list of household consumption and expenditure surveys: of almost 90 surveys from more than 25 countries, more than 75 use recall.

[2] The scale of the survey in Iraq precludes the use of individual diaries, which is deemed the most reliable survey instrument in Beegle et al. (2012). This difference is irrelevant here, as we do not rely on the assumption that diaries are error-free as they do. The experiment is interesting because the cheaper option is randomized (recall questions) while the expensive option (diaries) remains at baseline for all households surveyed.

based on IHSES diaries for 2007.[3]

Compared to prior work, our innovation is in the use of differences between diary and recall measurements for the same household, combined with differences in the reports of households assigned to different experimental groups. This setting allows us to tackle the identification problem in a much more general fashion compared to previous studies. In particular, we study the quality of reports from diaries and recall modules allowing for survey effects in both measurements. Our model is non-parametrically identified under mild conditions, discussed below. It yields identification of survey error distributions for recall and diary reports without assuming that errors from alternative collection modes are mutually independent, or independent of true consumption. These assumptions are rejected in our data and would be most likely violated in empirical settings (as conjectured by Aguiar and Bils, 2015). The distribution of true consumption underlying an household's diary and recall measurements is also non-parametrically identified. We can therefore assess the extent to which diaries deliver mean and distributional indicators close to those that would be computed using true consumption. The availability of repeated diary-recall measurements of food spending for the same household is not unusual, the CEX Diary Survey in the United States and the Canadian Food Expenditure Survey being notable examples. Our approach therefore extends to a number of contexts beyond the specific case study.

Among the conditions needed for non-parametric identification discussed in Section 5, two play a fundamental role. First, we assume that the average spending across diary households coincides with the average consumption of those households. We consider this a mild assumption. It is consistent with (and considerably weaker than) all papers that have investigated how consumption expenditures are affected by the collection mode. Also, its theoretical case can be made using a simple model of purchasing behavior (as in Meghir and Robin, 1992), as we discuss below. Second, we require variability in survey errors induced by an exogenous change in latent consumption. This embeds a standard exclusion restriction in the econometric literature on measurement errors, as discussed in Hu and Schennach (2008) and Chen et al. (2011) among many others. We elicit exogenous variability through the survey design. Households of participating enumeration areas of each district were randomized to interviews in different months of the year-long survey. True food consumption is seasonal in Iraq, as in most developing countries, due to variations in prices, the availability of certain food items, and important festivals and holidays.[4] We document the "first stage" requirement of within-year variability in consumption by looking at average spending depicted by diaries over survey months. The exclusion restriction here requires that the same household, if randomized to different survey months, would have had different survey effects only because of seasonal differences in her underlying consumption. This assumption would be violated if, for example, the same amount of consumption was reported with systematically higher or lower accuracy across months of the same year. One example of this could be seasonal variability in work or leisure which effects the willingness to engage in the survey. We discuss this and other threats to identification in the empirical section.

Our first finding is that diary and recall reports from the same respondent are not rank preserving, meaning that they do not order the household identically in the population distributions. We rule out simple explanations for this result. For example, we show that answering the additional recall module does not affect the accuracy of survey reports: households randomized to, and households excluded from the recall module, have the same spending in diaries. Important departures from rank invariance are found across the support of diary and recall measurements, suggesting that the survey effects are unlikely to arise only because of households with specific demographics and income. These results imply, for example, that similar households would end up above or below the poverty line depending on the survey mode employed. This is a worrisome conclusion for poverty and inequality measurement

[3] Fieldwork and resource constraints meant that this list was consolidated to include 20 such groups.

[4] Indeed, one of the important rationales for year-long surveys is to be able to capture seasonality in consumption and livelihoods. In Iraq, seasonality in consumption is experienced due to differences in availability of food items and importantly, due to month long observations of Ramadan, as well as two Eid holidays.

and means-testing. The sizable departure from rank invariance is also worrying because this is among the weakest assumptions needed to deal with measurement error in empirical work.

We use our identification strategy to establish the second finding – that little empirical support exists for claiming that diaries outperform recall measurements. Our results show that which survey method works best ultimately depends on the policy question under consideration. In general, however, the assumption of error-free diaries is misplaced: the variance of the difference between diary spending and true consumption can be substantial. Smaller values of this difference are found for households with the highest consumption, a fact consistent with survey errors decreasing in literacy. For example, the chance of extreme under-reporting consumption, less than half of the true value, using a diary is 16.3% and 12.9% for an household at the first and third consumption quartile, respectively. Attributing a value of household consumption at least twice as large as the true one is 5.3% and 4.8% for respondents at the first and third quartile, respectively. We show that these effects, if combined across all respondents, yield poverty statistics which are more reliable using recall data than diaries. We take stock of the frequency of purchases as the main channel to explain the mismeasurement in diaries, and estimate our model for components of food consumption that should not be subject to the frequency mechanism. We use item-level entries in individual diaries for the 2007 and 2012 rounds of the IHSES to flag those components of food that almost all households recalled having consumed in the week prior to the diary, and that are purchased frequently during the diary interview. Consistently with our story, we find much smaller (yet non-zero) diary errors for items for which *spending* is most likely equivalent to *consumption* in the interview week.

We then document our third finding – that recall errors are far from being classical in form, and over-reporting of true consumption is more likely than under-reporting. This finding is consistent with telescoping effects, but also with more general cognitive errors of respondents in the computation of the market value of their consumption. We find that the quality of reports improves with the value of true consumption, suggesting that recall errors may be comparatively less important in households with higher permanent income and human capital. Our analysis also shows that recall errors are more contained for those food components purchased more frequently.

Our fourth and final finding is that there isn't any loss in accuracy from using recall questions compared to the higher costs of using a diary. In the light of the results above, the question remains of whether consumption modules in large-scale household surveys should employ diaries or recall questions. Our empirical investigation shows how repeated measurements using both modes of collection for a subset of randomly selected households can be employed to elicit a superior measure of consumption, which is in the spirit of prior work by Browning and Crossley (2009). The choice of the most appropriate survey mode to use when measuring household consumption can be seen as the solution to a decision problem in which each household is assigned to a diary or a recall module so to minimize the overall impact of survey errors. The last part of our empirical analysis studies the optimal allocation that would make inequality and poverty measurements as close as possible to their true values. We do not find evidence to support the claim that the accuracy of policy conclusions is compromised by using recall data.

Why should our findings be of relevance to statistical agencies, policymakers and researchers? Faced with high costs of implementation, household fatigue, the need for enumerator effort and the risks of enumerator shirking, developing countries have been increasingly moved towards recall modes of data collection for food in their national household surveys.[5] However, there is little evidence on how to manage this transition, with most countries accepting a break in the series on poverty, food security and welfare aggregates. Our findings can inform the design of a protocol for the harmonization of time series resulting from such a transition and to prevent such a break in series. In addition, our findings

[5] Iraq is part of a broader trend in the Middle East and North Africa, where national statistics agencies are actively considering a transition from diary to recall. For example, faced with increasing non-response rates and high implementation and supervision costs, Jordan has most recently shifted from diary to recall for food in its Household Income and Expenditure Survey; Palestine and Lebanon are both considering it.

identify distribution-wide implications such as errors with respect to the true welfare ranking. These are directly relevant to policies that use proxy means tests derived from expenditure or consumption data to prioritize households for eligibility into social safety net programs. Finally, our findings have implications for the design of food consumption modules to minimize survey errors. Indeed, the design of food consumption modules is already evolving to explicitly measure frequency of purchase, and anchor questions on consumption rather than acquisitions (see for instance, the 2016 Barbados Survey of Living Conditions and the 2017/18 Jordan Household Income and Expenditure Survey). Our findings can provide the theoretical and empirical rationale for further refining food consumption modules.

The rest of the paper is organized as follows. The next section presents the general formulation of the problem. Section 3 describes the institutional background and our data. Section 4 directs the spotlight to preliminary descriptives and graphs about the validity of the diary-recall experiment. Section 5 presents the conditions for non-parametric identification of the survey effect distributions along with the distribution of true, latent consumption. Empirical specifications used for estimation are discussed as well, along with the exogenous variability used for identification. This section also reviews possible threats to validity in our research design. Section 6 presents our results on the nature and consequences of survey errors. Section 7 looks at the optimal combination of diary and recall respondents that would yield data with minimum distance from true consumption. Section 8 concludes.

2 General Formulation of the Problem

2.1 Quantity of Interest

Let Y^d and Y^r be the measurements for food reported using diaries or recall questions, respectively. These are possibly error-ridden indicators of true consumption Y^*. Household is the statistical unit of analysis. Interest lies in the distribution of true consumption, $F_{Y^*}[y]$, which is the quantity needed to retrieve key location and inequality functionals used for policy analyses. For example, as consumption is non-negative we have:

$$\begin{aligned} E[Y^*] &= \int_0^\infty (1 - F_{Y^*}[y])dy, \\ G[Y^*] &= 1 - \frac{1}{E[Y^*]} \int_0^\infty (1 - F_{Y^*}[y])^2 dy, \end{aligned}$$

where the last expression is the Gini coefficient. Household demographics and area characteristics are also available, the conditioning on which is left implicit throughout.

Small contamination from classical errors would raise the density of Y^* where it is convex and flatten it where it is concave, increasing dispersion (Chesher, 1991). It follows that even the simplest form of mismeasurement in Y^d or Y^r has implications for welfare analysis, for example by increasing the Gini coefficient for the distribution of error contaminated consumption measurement (Chesher and Schluter, 2002). The effects of non-classical errors on $F_{Y^*}[y]$ an its functionals, such as the Gini coefficient or the poverty rate, are difficult to sign in general and will depend on features of the true distribution as well as that of measurement errors. Non-classical errors in consumption measurements are ubiquitous in empirical work, as we explain below and is discussed by Bound et al. (2001).

Randomizing households to different modes of consumption measurement is not sufficient to retrieve the distribution of recall errors, even when diaries are error-free. This is a largely overlooked problem in the literature, which has used experiments as the smoking gun to learn about reporting errors. Let D be the an indicator for household reports obtained from diary interviews and $Y = Y^r + D(Y^d - Y^r)$. Randomization combined with the assumption that diaries are error-free imply $F_Y[y|D = 1] = F_{Y^*}[y]$. The difference between $F_Y[y|D = 1]$ and the distribution obtained using recall questions, $F_Y[y|D = 0]$, conveys information on recall errors. However, as randomization reveals only one potential measurement, recall errors are not identified at the household level nor is their distribution across households.[6]

2.2 Origin of the Measurement Equations

The implications of using alternative methods of collecting consumption data through household surveys are the object of a voluminous literature spanning both developed and developing countries (see Deaton and Zaidi, 2002, and Carroll et al., 2015, for reviews). While there is consensus on the sensitivity of consumption estimates to the collection mode, the lack of validation data generally precludes definite conclusions regarding the most accurate method. Validation studies are scarce, and often data quality is inferred from ingenious comparisons with other sources that may have other measurement problems (like the national accounts). Surveys based on diaries are usually considered the "gold standard" in developing countries for measuring household expenditures. However, a careful review of the literature suggests that this conclusion should be taken with a grain of salt.

Recall Errors

Why should one expect errors in recalled consumption? A vast literature has investigated the drawbacks of, and the errors from using recall interviews (see the review by Lyberg and Kasprzyk, 2011). In most LSMS surveys households must recall the quantities of a pre-selected list of food items consumed over a fixed reference period, and their associated market value. In our application, for example, the recall period are the 7 days ending with the interview (Beegle et al., 2012, and Crossley and Winter, 2014, discuss the effects of lengthening or shortening this time horizon). Information here is collected in two steps. The household 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 or received as gifts or 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. This procedure is the most common way to collect recall food consumption information in household surveys in developing countries.[7]

This questionnaire design has important implications for understanding what to expect about the properties of reported consumption. For starters, recall interviews ask the value of consumption, not spending. Limited cognitive abilities and the difficulty of recalling the timing of consumption and the associated market expenditures where relevant may challenge the computation of respondents. 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 underlying true value Y^*. Experimental evidence in psychology research supports this interpretation (see Menon, 1993, and Comerford et al., 2009). Rounding errors in providing the quantities consumed or telescoping (the act of recalling consumption occurred over a longer period

[6] The weaker assumption that the distribution of diary errors is centered at zero yields identification of the mean of Y_i^*. Indirect evidence on the properties of recall errors can be obtained by comparing the diary and recall reports of clusters of households, such as those living in the same village (see Gibson et al., 2014, for an example).

[7] Typically, information on quantity consumed (but not their associated market value) is used to estimate a range of food security measures including caloric deficiency, nutritional deficiency and dietary diversity. Often, the estimation of these requires additional questions on whether certain aggregated food groups were consumed by the household and the frequency of consumption.

of time into the reference period; Neter and Waksberg, 1964) were found to be important factors in the computation. The errors resulting from this cognitive process don't have an obvious sign, but we know that consumption measurements vary with features of the recall interview (as revealed by the experiment in Beegle et al., 2012).

The textbook assumption of classical measurement error is therefore violated, 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 centred at zero, not correlated with Y_i^r but correlated with true consumption Y_i^*. 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, for examples of such reporting errors).

Diary Errors

Diaries are self-administered forms in which households must record all expenditures for the duration of a defined period in survey, including market purchases and estimated market values of consumption from non-market sources. Reports are not exempt from errors: the effects of fatigue on accuracy have been known to researchers for long time (as documented in Silberstein and Scott, 2011). In our example and most empirical applications, the quantity Y^d is defined as total spending on food over the duration of the interview.[8] It is common practice that households are assisted by frequent visits of enumerators during the recording process. For instance, in the Iraq survey the diary is dropped off on day 1, when the enumerators gives instructions to the household on filling out the diary, starting from the next day. This is followed by 4 additional visits in the next 8 days to ensure that households have filled out the information in time and are following instructions. It is generally presumed that this design yields the most accurate data (see Brzozowski et al., 2017, for discussion).

The assumption that diary records are error-free is not enough, alone, to elevate them to benchmark against recalled consumption. When the underlying value of consumption Y^* is of interest, diary errors may arise because of differences between *consumption* and *spending*. This aspect of the comparison has been largely overlooked in empirical work. For instance, in the 2012 Iraq survey discussed below, in diaries households report purchased quantities over the 7 day reference period, while in recall households report consumed quantities over the 7 day reference period. The same setting is found in the Tanzania experiment by Beegle et al. (2012). Some products, such as wheat, rice, and oil can be purchased in bulk and stored. If the frequency of purchases varies across households depending on their access to markets, preferences or liquidity constrains, diary errors are likely to vary with Y^*. This mechanically invalidates the textbook assumption of classical errors. Also, the frequency of purchases has more pronounced effects the shorter the length of the diary period, which is 7 days in many countries. For instance, in the Iraq data, among households who report both diary and recall measurements, the share of households with positive consumption reports for staples such as wheat and rice is much higher than the share of households with positive purchase reports, as we show below. Moreover, the variance of quantities purchased reported in diaries is much higher than that of consumption from diaries, suggesting the presence of infrequent and bulk purchases.

The theoretical case for differences between consumption and spending can be made with the aid of a simple model of purchasing behaviour. Assume that the household has a target monetary amount of consumption τY^* over τ weeks, and that equally-sized purchases are made over this period. Consumption is smooth over time, but may differ from expenditure in the diary week. Let N denote

[8] Instructions for the Iraq diary module ask households to record their "daily expenditure for a period of 7 days from the day following the day you received the notebook", including daily expenditure on food and drink as well as recurrent non-food expenditures, followed by instructions on how to report expenditures for non-purchased items. The quantity Y^d may include market valuations of consumption, as in recall, but only for the small quantities of food derived from non-market sources such as self-production, barter or gifts. In practice, however, market purchases account for almost all of the spending in our case study, as it is often the case in other empirical applications (even in developing countries).

the number of purchases recorded in the diary week, and let N^* be the expected number of purchases in a typical week. This setting implies that Y^*/N^* is the amount spent in each purchase, and:

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

Diary errors arise from differences between household consumption Y^* (the quantity of interest) and household spending Y^d (measured from the diary). It is clear that this error may be non-zero even if all expenditures are listed correctly in the diary. What are the implications of this problem for empirical work?

The take-away message is that there is little theoretical support for considering diaries the most reliable source to measure consumption poverty. Additional assumptions are needed in general. In our model, for example, the assumption $E[N|N^* = n^*, Y^* = y^*] = n^*$ implies $E[Y^d|Y^* = y^*] = y^*$. The use of diaries to measure consumption averages can be grounded on this implication. However, our model also implies that the variance of Y^d can be quite different from the variance of Y^* if the frequency of purchases is low.[9] This interpretation may explain the differences between inequality indicators obtained using diary and recall data (see Beegle et al., 2012, Battistin and Padula, 2016, and Coibion et al., 2017).

2.3 Aggregation over items

Our empirical analysis uses a measure of consumption computed by aggregating over a number of food items. Total food errors are defined by adjusting definitions above and considering I items:

$$\begin{aligned} Y^d &= \sum_i Y_i^d, \\ Y^r &= \sum_i Y_i^r. \end{aligned}$$

Aggregation exhacerbates the non-classical properties of recall errors, and the fact that they will likely correlate with true consumption. Besides, the assumption that diaries can be used as benchmark is more nuanced than it may appear at first. Our model yields the following expression for diary errors:

$$Y^d = Y^* \left(\sum_i b_i^* \frac{N_i}{N_i^*} \right), \tag{1}$$

where $Y^* \equiv \sum_i Y_i^*$ is food consumption and $b_i^* = \frac{Y_i^*}{Y^*}$ is the budget share devoted to item $i = 1, \ldots, I$. The distortive effects of errors on the measurement of consumption functionals (e.g., the share of consumption-poor households) are difficult to sign. If participation in the survey affects the purchasing behaviour of respondents, diary errors will not be centred at zero in general. It is well known that

[9] The model implies:

$$Var(Y^d) = Var\left(Y^*\right) + E\left[\left(\frac{Y^*}{N^*} \right)^2 Var(N|N^* = n^*) \right],$$

meaning that the variance of diary expenditure may be considerably larger than the variance of true consumption. For example, assuming no heterogeneity in frequency behaviour with purchases occuring at random and independently at a rate of one per week ($N_i^* = 1$ for all households), the variance of true consumption would be overstated by a factor larger than two:

$$Var(Y^d) = 2Var(Y^*) + E[Y^{*2}].$$

The model can be generalized to allow for purchases of unequal size and/or random errors in reporting. The literature on (in-)frequency of purchases has a longstanding tradition (e.g., see Meghir and Robin, 1992).

households tend to spend more – or report more spending – at the beginning of the diary period (in the first day in particular; see Lyberg and Kasprzyk, 2011, and Silberstein and Scott, 2011). This evidence is consistent with behavioral responses causing N_i to depart from the usual N_i^*. For example, households may purposefully decide to move to the diary week expenditures that would be otherwise delayed to a later time. Such manipulation of purchases will likely depend on Y^*.

2.4 Related literature

Our work connects with a large literature investigating whether increased income and earnings inequality in the United States has been tracked by consumption inequality (see Aguiar and Bils, 2015, and Attanasio and Pistaferri, 2016, for examples). Many studies have addressed this question using data from the CEX Interview Survey or the PSID, which measure spending based on recall questions (see Krueger and Perri, 2006, Blundell et al., 2008, Attanasio and Pistaferri, 2014, and Blundell et al., 2016). As cross-sectional inequality has risen more rapidly in the CEX Diary Survey compared to what measured in recall data (Battistin, 2003), the question arises of whether measurement problems can explain this puzzle. The high-frequency of diary data is a possible candidate: changes over time in the frequency of purchases, combined with a rise in their size, may trigger higher variability in measured spending almost mechanically. The frequency of purchases channel we address here echoes the explanation offered by Coibion et al. (2017) using diary data for the United States.

Our results also connect with the literature considering consumption dispersion between demographic groups which proxy permanent income (such as education; see Attanasio and Davis, 1996). Here the empirical question is to compare the distance in expenditure between households or individuals in the top and in the bottom groups, relative to the corresponding differences in income, over time. This between-group statistic is not affected by changes in the frequency of purchases. However, our results below show that, with recall data, consumption averages are effected by errors that depend on latent consumption and therefore income. Differential recall effects across income groups may affect the size of the between-group consumption gradient. Our findings can inform how to adjust for this bias in empirical work.

We also connect with the literature on proxy-based poverty measurement.

3 Background and Data

3.1 The Iraq Household and Socio-Economic Expenditure Survey

Iraq is an upper middle income, resource-rich, fragile, and conflict-affected country. It faces development challenges in line with far poorer/lower income countries. The median education level in 2012 was primary schooling or below. Infant mortality rates remain below the norm for similar income countries, a third of Iraqi children aged 0-5 are stunted, only 58 percent of adult males of working age (15-64) are employed, and only one in five women of working age participate in the labor market. After the fall of Saddam Hussein's regime, the country ended a period of relative isolation and adopted international standards for measuring poverty and other socio-economic indicators. A new Living Standards Measurement Survey (LSMS) started in 2007, the Iraq Household and Socio-Economic Survey (IHSES).

Our primary source of data is the second round of the IHSES, implemented in 2012, which was designed to be comparable to the 2007 round. The IHSES provides data for official statistics at the national and sub-national level on poverty, food security, income, labor market outcomes, health and education. The survey is also the basis for detailed information on the household consumption component into national accounts, as well as national 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 includes special modules to fill critical data gaps (like anthropometrics and time use).

The IHSES 2012 had an intended sample size of 25,488 households and a final sample size of 24,944 households. The population was stratified on 119 districts (qadahs). Within each district, 24 census enumeration areas (EAs) were randomly selected and 9 households sampled with equal probability within each EA. Teams consisting of three interviewers and one supervisor 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.[10] It follows that, by design, households in one EA had an equal probability of being interviewed in any of the survey months.

3.2 Food Consumption Measurements

Food measurements in the IHSES 2012 were collected using a 7 day diary, which was handed to the household during the first of five visits. Entries in the diary were recorded by the head of household or the most informed respondent, and the diary was assisted. Enumerators visited each household five times over 9 days (every alternate day) checking if expenditures were entered correctly, clarifying any questions, and entering data already recorded in the diary since the last visit into computers. Instructions emphasized that the household should record daily expenditures on repetitive or recurrent food and non-food commodities, and meals taken outside of the household starting from the next day.

One-third of households within each EA were also administered a recall module on food consumption during the first visit. This randomly selected group of roughly 8,000 households constitutes our "treatment" sample.[11] It follows that, by design, two measures (diary and recall) are available for household 3, 6 and 9 in each EA, and one measure (diary) is available for all remaining households. The recall module covered the 7 days prior to the enumerator's first visit to the household. 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*". The availability of multiple measurements, as well as the scale of the experiment, mark a departure from past work. However, unlike Beegle et al. (2012), we do not have individual diaries or varying recall periods for food consumption. Our work shows instead the potential for household surveys of eliciting repeated measurements from the same respondent, which can be used to learn about reporting errors and the underlying consumption patterns (this is in the spirit of prior work by Browning and Crossley, 2009).

Infrequent or bulk purchases are less likely to be recorded in diary expenditures relative to recalled consumption. The divergence between purchase and consumption frequency is pervasive in the developed and the developing world. Take, for instance, wheat or rice, which are consumed daily but are purchased at much lower frequency. More perishable items such as fresh vegetables and fruit may have higher purchase frequency but are still likely to be consumed daily when available and affordable. In developing countries where food is purchased at weekly or bi-weekly markets, or where access to markets is restricted or costly, consumption of a broad range of food items is systematically more frequent than food purchases. Therefore, while diaries may still measure consumption on average, they will have higher variance and different distribution. This difference in the distribution has implications for which collection mode, diary or recall, yields a better measure of the true consumption

[10] 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 in each wave, the teams alternated the EA visited each day.

[11] Three different modules were administered in mutually exclusive and exhaustive subsamples of 3 households per EA: anthropometrics, time-use and food consumption by recall.

distribution. That being said, the level of enumerator effort in Iraq to ensure diary entries were correctly and regularly recorded is high. In addition, there are many reasons to believe that recall modes of collection exacerbate certain errors. These include the possibility that households forget to record the consumption of certain items or misreport the value of consumption due to telescoping error or because values are roughly estimated based on rules of thumb or backcasting. These concerns may be particularly salient in a context such as Iraq where many households have relatively low educational attainment. There are no incentives for households to report specific expenditure 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.

We compare expenditures in diaries with the market value of recalled consumption. Iraqi households receive in-kind subsidized food through the PDS, or food rations, at negligible cost.[12] Information about PDS items is collected in two separate modules: the rations module and a 7 day diary for food purchases. The former collects information about the quantity of ration items received, consumed, bartered, sold or given away by the household during the last 30 days. The diary records all market purchases of food including ration items over the last 7 days (expenditures and quantities). In effect, the diary records market transactions of households to purchase ration items over and above their monthly allocation, and these transactions are rare and small in magnitude. Because our focus is a comparison of diary expenditures with recalled consumption, the estimated value of rations consumed reported in the rations module is excluded from our analysis. Both diary and recall modes in Iraq include household estimations of the market value of food items consumed that are not acquired from the market such as self-production and gifts. Such valuations are included in both diary and recall measurements. Finally, as the recall module included a selected list of 20 aggregated food items, these were matched carefully to their counterparts in the diary (diary food items are quite disaggregated, and use the COICOP classification as we show in the Appendix). Our recall module bears the closest resemblance to the subset list of the 17 most important food items considered in the Beegle et al. (2012) Tanzania experiment.

The transition to recall being considered in Iraq is part of a broader trend across the developing world, where countries are shifting to recall for lower implementation and supervision costs, and greater ease in soliciting responses. Non-food expenditures in Iraq and other parts of the developing world are already measured using different recall periods (depending on the frequency of purchase), so this experiment in Iraq was limited to food expenditure. The objective of testing an alternative was to provide information to guide an eventual transition to recall as a cheaper and easier way of collecting this data, and of potentially allowing some kind of translation of trends from two otherwise non-comparable series. Jordan has most recently shifted from a 15-day diary to a recall module for food for the first time. In Jordan's case, a full revision of the questionnaire implied a break in series, but other countries such as Iraq are moving more cautiously to consider the implications on series comparability. In this context, the Iraq experiment provides a unique opportunity to estimate the size and magnitude of errors under both modes relative to the true distribution, characterize implications for change in ranks, especially for the bottom of the distribution, and to quantify the differences in the two series for welfare measurement.

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

4 Descriptives and Graphical Analysis

4.1 Covariate Balance

We begin by documenting in Table 1 balance across households with and without randomly assigned recall modules. Specifically, this table shows regression-adjusted treatment-control differences from models that control for strata (EAs) used in the randomization design. The working sample here is obtained by pooling households in the baseline and treatment samples. All variables are well-balanced across groups, as can be seen in the small and insignificant coefficient estimates.

Randomization to the recall module does not affect the reporting of expenditures in the diary, as shown by the coefficients in Table 2. Respondents might adjust their diary entries to reconcile aggregates with those in the recall module, as in bounded interviews (see Neter and Waksberg, 1964). Results in column 4 of the table rule out this possibility. Reported in Panel A are coefficients from plain and quantile regressions on a treatment dummy and strata controls using diary reports for both the baseline and the treatment samples. Households randomized to the recall module have the same distribution of diary consumption as those in the baseline sample. Panel B tests for differences in disperison considering the standard deviation of logged reports and the Gini coefficient. We conclude that participation in the recall module does not affect the reporting of diary entries.

Recall data yield higher consumption values than diaries. This can be seen from column 5 of Table 2, which presents coefficients from the same regressions in column 4 but using, on the left hand side, diary reports for the baseline sample and recall reports for the treatment sample. Larger values for 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, Bee et al., 2013, and Battistin and Padula, 2016). This result is also consistent with smaller diary-recall differences for recall periods langer 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 telescoping of consumption.

A visual inspection of diary-recall differences across the support of the consumption distribution is in Figure 4, below. The take-away message here is that interview effects are heterogeneous across households: differences between distributions are not a simple location shift, but picture different effects across percentiles. This evidence weighs against mean diary-recall differences as the most interesting quantity to consider. It also suggests that the interview mode must bring along subtly nuanced effects on measured consumption inequality which cannot be understood by a simple treatment-control comparison. We dig deeper into this idea next.

4.2 Ranks

We use non-parametric plots to compare the rank of the same household in diary and recall distributions. The analysis is carried out using the treatment sample, for which both measurements are available.

Diary and recall reports from the same respondent do not order her household identically in the population, a fact shown in Figure 1. Here the horizontal axis shows household ranks computed from diaries. On the vertical axis, the same households are ranked in the recall distribution. The continuous line 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 using CEX data for the United States. Rank invariance is among the weakest conditions required to deal with measurement error in empirical work.

Differences between recall and diary measurements are not mechanically explained by household income. This can be seen from 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 income distribution are, on average, in the second quintile of the consumption distribution. Households in the top income quintile can be, on average, as low as in the third consumption quintile. This relationship does not change with the survey mode employed, suggesting that diary-recall differences can't be explained by differential errors at different points of the income distribution.

4.3 Anatomy of Diary Errors

Sharp differences across food items emerge as to household reports of consumption and spending. This can be seen from Panel A of Figure 3, where we compare 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). Using the randomization design, the former quantity is computed by pooling diary data from the baseline and treatment samples. The vertical line around bars shows how the share with positive spending varies across survey months. For example, 40% of households report positive spending on rice in the diary week, and this number varies between 30% and 55% depending on the survey month. On the other hand, almost all households recall positive consumption of rice in the week preceeding the interview. The consumption pattern across items mirrors the spending pattern quite closely. Consumption is however more likely than spending, and significantly more so for items at the left-hand-side of the figure.

Frequency of purchases is the most likely explanation for this finding. Households are between two and three times more likely to report consumption of storable items than spending on these items in the diary week. This difference drops to zero for perishable items, like meat or fish.[13] Calculations not reported here show that almost one-third of consumption comes from items for which the difference between consumption and spending is the largest. The take-away message is that a large share of household consumption in Iraq comes from storable items. If these items are purchased less frequently and in bulks, large values of N/N^* in (1) may mechanically explain diary-recall differences in the data. The same channel has been suggested by Coibion et al. (2017) to explain differences in inequality measurements for the United States.

The frequency effects on survey reports apply to respondents from all socio-economic backgrounds. Panel B of Figure 3 replicates the analysis above stratifying on quartiles of household income. Not surprisingly, the likelihood of spending in the diary week grows with the income quartile. However, large differences between consumption and spending remain for storable items across income quartiles.

5 Non-Parametric Identification and Estimation Strategy

We consider a setting in which the reports Y^d and Y^r are observed for the same household. This is the case for the treatment sample in our case study, although the availability of repeated measurements for Y^* is not unique to Iraq. The Interview component of the CEX (Battistin, 2003) and the Canadian Food Expenditure Survey (Brzozowski et al., 2017) are notable examples of surveys with repeated diary and recall modules on food spending. Our strategy for understanding the role of survey errors in the IHSES therefore extends to data widely used in empirical research for Canada and the United States.

The availability of repeated measurements ensures identification of the distribution of true consumption and errors if diary and recall errors are both mean zero, mutually independent and indepen-

[13] A caveat for the interpretation of these differences is the mis-alignment of diary and recall interview periods. The latter period refers to consumption occurred in the week ending with the survey, while the former period to the following week. We believe that the pattern across items could be hardly explained solely through this channel.

dent of true consumption Y^* (Kotlarski, 1967). Yet, we have discussed reasons why the assumption that errors are independent of Y^* is likely violated in most empirical settings. Instrumental variation combined with the availability of repeated measurements solves for this limitation, ensuring non-parametric identification of the distributions of Y^* and measurement errors in both diary and recall records. This is the approach we take here, which builds on prior results in Hu and Schennach (2008).

Our identification strategy doesn't hinge on any parametric assumption about the underlying distribution of consumption or error. Moreover, we do not impose that diaries are error-free and allow diary and recall errors to depend on the true underlying consumption Y^*. These features mark an important departure from prior work on consumption measurement, and allow us to test (and reject) many assumptions made in past research.

5.1 Assumptions

We assume throughout that latent consumption and its measurements (Y^*, 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 true or reported 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 normally distributed consumption is made in Battistin et al. (2009).

Identification stems from an exclusion restriction that brings in the picture a particular type of instrumental variation.

Assumption 1 *(Exclusion Restriction) There exists a continuous variable Z such that:*

$$(Y^r, Y^d) \perp Z | Y^*. \tag{2}$$

This is a standard restriction in the econometrics literature on measurement error (see Chen et al., 2011, and Schennach, 2013, for examples, and Assumption 2 in Hu and Schennach, 2008). 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. Importantly, the variable Z can be arbitrarily correlated with errors in both diary and recall data. A more general interpretation follows from the requirement of conditional independence: knowledge of Z must not yield any more information on survey measurements than Y^* would otherwise provide.

Our empirical investigation employs randomness in the timing of interviews to define Z. We assume that respondents randomized to interviews in different waves of the year-long survey will report with different errors only because of within-year variation in household consumption. This condition would be violated if the accuracy of diary entries or the cognitive abilities while reporting the same amount of consumption varied systematically over months of the survey year. As the survey was carried out over 24 consecutive two-week long waves throughout 2012, we will exploit random assignment of interviews to different days of the year. Implicit here is a “first stage” condition requiring enough within-year variation in the underlying household consumption Y^*. This can be due to seasonality in access to different food items, as we discussed above.

Diary and recall errors can depend on each other, and with true consumption. This is a likely possibility in empirical work, and is contemplated by Assumption 1. However, we assume that the correlation between diary and recall errors is channeled through true consumption.

Assumption 2 *(Conditional Independence) The repeated measurements are conditionally independent given true consumption:*

$$Y^r \perp Y^d | Y^*. \tag{3}$$

This setting is flexible enough to study very general forms of misreporting and to relax the assumption of independent errors which is routinely made in the literature (Assumption 1 and Assumption 2 here imply Assumption 2 in Hu and Schennach, 2008). Our Assumption 2 is violated if, for example, the respondent's tendency to misreport their recall consumption at the beginning of the interview affects the diary accuracy later in the survey. We believe this is unlikely to be a threat to identification in our setting, as differences between N_i and N_i^* in (1) should not depend systematically on recall errors for reasons other than Y^*. The evidence in Table 2 weighs against this channel, as we don't find evidence that participation in the recall module affects spending during the diary week.

Our model of purchasing behaviour in Section 2 motivates the following additional restriction.

Assumption 3 *(Mean Zero Errors) The distribution of diary errors is centered at zero:*

$$E[Y^d - Y^*|Y^* = y^*] = 0.$$

This assumption implies that spending records obtained with assisted diaries yield an unbiased measure of household consumption, $E[Y^d] = E[Y^*]$. It is considerably weaker than assuming that diaries provide an error-free measurement of weekly consumption, which is the traditional war-horse in most of the empirical work we are aware of. Although the assumption is stated here with reference to the conditional mean of diary errors, Hu and Schennach (2008) show that identification is still achieved if any measure of location or some other functional of the conditional distribution of Y^d given Y^* is known. For example, if respondents are equally likely to overreport or underreport the truth, but not by the same amount, then the median of measurement error (defined as $Y^d - Y^*$) would be zero. Similarly, if reporting values close to the truth is more likely than reporting any value far from the truth, then the mode of the measurement errror would be zero. Our choice for Assumption 3 follows naturally from the properties of measurement error arising from (1) and ensures ensures a computationally simple estimation. Also, it is general enough to be valid even when the frequency of purchases is manipulated by respondents.[14]

Two additional formal conditions are needed for identification, which are unlikely a limitation in our setting. First, true consumption Y^* must have a causal effect on the distribution of recalled consumption Y^r (this is equivalent to Assumption 4 in Hu and Schennach, 2008).

Assumption 4 *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 $y_1^* \neq y_2^*$.

This is a mild restriction, which is satisfied if $E[Y^r|Y^*]$ increases monotonically in Y^*but it is much weaker than that. We also require sufficient variability in the support of the conditional distributions of Y^d given Y^* and of Y^* given Z:

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

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

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

[14] The assumption:

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

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.

Completeness assumptions are common in the literature on non-parametric identification with instrumental variables, and are 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 $h(Y^d) = 0$ (Assumption 5 is the equivalent of Assumption 3 in Hu and Schennach, 2008). 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 essentially a generalization of the rank assumption on the instrument Z to the non-parametric setting.

Assumptions 1-5 imply:

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

and that there exists a unique choice of distributions on the right-hand side that generates the observable distribution on the left-hand side. The identification result does not hinge on any parametric assumption, and follows from adapting the results in Hu and Schennach (2008) to the case of repeated measurements.

5.2 Empirical Specifications

The case for the empirical relevance of the timing of interviews for consumption is made in Table 3. Panel B here shows regression of Y^d on the interview bimester controlling for a full set of district effects. Bimesters are defined to reflect the agricultural seasons (lean or off-season, harvest season and post-harvest season), as explained in the table footnote. Under our maintained Assumption 3, the expected value of Y^d in each bimester equals the expected value of the unobserved Y^*. The results show high seasonal variability in total spending in the year-long survey, with particularly high expenditures in the first and fourth bimesters. Panel A shows similar regressions where we use several household characteristics. As the day of the interview is randomized across EAs of the same district, we expect no effects. This expectation is borne out by column 6 of the table, which reports p-values for F-tests of the joint significance of the coefficents on the five bimesters. All p-values are remarkably large. Building on the day of the interview randomization, we construct an instrument Z as the predicted value of a non-parametric local polynomial regression of Y^d on the day of the interview after netting off district fixed effects.

The empirical challenges in estimating unknown densities in (4) are addressed by using a flexible functional form for the conditional densities on the right hand side of the expression. Building on the invariance property of the likelihood function, we consider log transformations for the quantities involved to ensure well-behaved distributions closer to normality. Following Barron and Sheu (1991), we consider densities that belong to a flexible class of exponential families. 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:

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

is a normalizing constant and $L_k(x)$ is the k-th Legendre polynomial defined on the interval $[-1, 1]$. Legendre polynomials are chosen here because of their numerical properties in this context (see Crain, 1974, and Crain, 1977). The quantity Δ_r is set to 4 in our analysis, as this ensures that the argument of $L_k(x)$ lies in the interior of the interval $[-1, 1]$ for any reasonable choice of $\log y^r$ and $\log y^*$. The smoothing parameter K_r determines the degree of departure from normality of the distribution in (5). The value $K_r = 2$ corresponds to imposing a normal distribution, while a value $K_r > 2$ allows for significant departures from this case. We choose the normal distribution as baseline since the observed marginal distribution of $\log y^r$ are close to normality; as such we expect a good approximation for the conditional densities of interest with a relatively small value of K_r. Still, we explore significant departures of K_r from 2 as we explain below. The dependence of the distribution on $\log y^*$ is modeled by imposing:

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

where the quantities δ_0^k and δ_1^k are set to 3 and 7, respectively, to ensure that the argument of $L_j(x)$ is in the interior of $[-1, 1]$. Our empirical investigation considers $J_r = 2$, thus allowing each parameter θ_k^r in (5) to depend on $\log y^*$via a quadratic polynomial.

The density of $\log Y^d$ given $\log Y^*$ is modelled is a similar manner, and depends on a vector of parameters $\boldsymbol{\beta}^d$:

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

where the values $\Delta_d = 4$, $\delta_0^d = 3$ and $\delta_1^d = 7$ are considered, and:

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

Finally, the density of $\log Y^*$ given $\log Z$:

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

depends on a vector of parameters $\boldsymbol{\beta}^y$, where:

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

and the values $\Delta_y = 8$, $\delta_0^y = 0$ and $\delta_1^y = 3$ are considered.

The unknown parameters $\boldsymbol{\beta} = (\beta_{10}^r, \ldots, \beta_{K_r J_r}^r, \beta_{10}^d, \ldots, \beta_{K_d J_d}^d, \beta_{10}^y, \ldots, \beta_{K_y J_y}^y)'$ are estimated via sieve-maximum-likelihood (see Grenander, 1981, and Geman and Hwang, 1982). More precisely, we maximize the pseudo-likelihood function obtained by substituting the unknown conditional densities in (4) with their approximations considered here:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \log f_{\log Y^r \log Y^d | \log Z}[\log y^r \log y^d | \log z; \boldsymbol{\beta}].$$

By allowing the smoothing parameters K_r, K_d, K_y to increase with sample size, this procedure yields non-parametric estimates of the conditional densities of interest. By inspecting the value of the maximized likelihood at incremental values of the smoothing parameters, we select $K_r = K_d = 6$ and $K_y = 3$ as little improvement in the likelihood is found at larger values. The error distributions, $f_{\log Y^r | \log Y^*}$ and $f_{\log Y^d | \log Y^*}$, are therefore estimated away from normality. The conditional distribution of logged consumption, $f_{\log Y^* | \log Z}$, as well is not normal, although the degree of departure from this benchmark is not pronounced. The latter finding is consistent with the theoretical motivation for having log-normal consumption (Battistin et al., 2009). The estimated conditional densities are then transformed to levels to obtain the conditional densities of interest, $f_{Y^r|Y^*}[y^r|y^*]$, $f_{Y^d|Y^*}[y^d|y^*]$ and $f_{Y^*|Z}[y^*|z]$, which are discussed in the next section.

6 Errors in Survey Reports

6.1 Diary and Recall Measurements

The presumption that diaries outperform recall data finds little empirical support. This can be seen from Figure 4, where diary and recall histograms are compared with the estimated distribution of true consumption, $f_{Y^*}(y)$ (the continuous line). A visual inspection suggests that the lower tail of the recall distribution is closer to that of true consumption, and that mismeasurement from diaries is more substantial at the bottom end. Although diary measurements are centred at true consumption (because of Assumption 3), the difference in the modes of true and diary distributions suggests that diary errors must have a thick lower tail. On the other hand, it is evident how recall errors may be substantially larger at the upper end of the consumption spectrum, where diary interviews do a better job. These conclusions are confirmed from the first three columns of Table 4, where different percentiles of empirical and estimated distributions are compared. The bottom three deciles of the recall distribution are closer to the true deciles than using diaries. The opposite holds for the remaining deciles in the table. The Gini coefficient from the diary distribution overstates the real value, and this difference is mitigated by using recall data.

These findings have important implications for the computation of poverty statistics. This can be seen from Panel A of Figure 5, where any point η on the horizontal axis is a possible value of consumption in the true distribution. Here the curve for diary is $P\left(Y^d \leq \eta\right) - P\left(Y^* \leq \eta\right)$, which is the measument effect on the proportion on households scoring below η when diaries are employed. 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 headcount rate is sensibly higher using diaries than using recall data. This can be seen by noticing that the share of households consuming less than $\eta = 8.75$ USD a week (the bottom 19.1% in the true distribution) is 14.5% in the recall distribution, and 30.1% using diaries. Similar calculations can be used to compute the poverty gap, which measures the average distance to the poverty line (where households above the line are given a distance of zero) is the largest for diaries, at 9.2%. The corresponding shortfall computed in the recall distribution is 3.9%, and should be compared with the value 3.4% obtained from $f_{Y^*}(y)$. Given the trade-offs involved, which survey method works best ultimately depends on the policy question. When overall poverty statistics are of interest, recall data yield more reliable figures than diary interviews. A similar conclusion applies to inequality measurements, as we will see soon.

The assumption of error-free diaries is an illusion. This can be seen in Panel A of Figure 6, where reported are error distributions for an hypothetical household at selected percentiles of Y^*. The quantity Y^d/Y^* on the horizontal axis lends itself to a simple interpretation: for example, a value of 1.5 here means that the household measurement is 50% larger than her true consumption. Because of Assumption 3, all distributions in this panel are centred at one (the distributions shown here and below are transformations of the distributions estimated as explained in the previous section). Despite being correct on average, we find that diary measurements understate consumption for most households. The chance of severely under-reporting consumption using a diary ($Y^d \leq 0.5Y^*$) is 16.3% and 12.9% for an household at the first and third quartile, respectively. The chance of attributing a value of household consumption at least twice as large as the real one ($Y^d \geq 2Y^*$) is 5.3% and 4.8% for respondents at the same two quartiles. Panel B of Figure 6 offers a visual representation of 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 an higher probability mass and the dashed line shows the conditional expectations (which are equal to one). The mode of these distributions is below one in all cases, meaning that underporting is the most likely outcome.

Errors in recalled consumption are far from being classical in form, with over-reporting being more likely than under-reporting. This can be seen from Panel A of Figure 7, which reports error distributions considering values of Y^r/Y^* on the horizontal axis. The contour plot in Panel B demonstrates that the conditional means and modes get closer to one as true consumption increases. For example, the averages of the distribution Y^r/Y^* for households in the 90th and 10th percentiles of Y^* are 1.18 and 1.41, respectively, as shown in column (1) of Table 5. The average error for the household with median consumption is 1.248. Recall errors are therefore negatively correlated with true consumption.[15] Moreover, the standard deviation of these distributions shrinks as true household consumption increases. The effects on the mean squared error (MSE) of conditional distributions can be seen in column (3) of Table 5. The simplest story seems most likely: higher consumption is correlated with human capital and better cognitive abilities, and the ability to compute more reliable recall measurements. Poorer households tend to overreport consumption. The chance of severely under-reporting consumption using recall questions ($Y^r \leq 0.5Y^*$) is 3.2% and 1.4% for households at the first and third quartile, respectively, while the probability of reporting a value of household consumption at least twice as large as the real one ($Y^r \geq 2Y^*$) is 12.7% and 5.5% for respondents at the same two quartiles. The time pressure argument would suggest that those with higher incomes and less leisure should be less likely to respond to surveys. We find the opposite pattern.

Diary errors present important similarities with classical errors in measurement. We notice that error densities in Panel B of Figure 6 grow slightly bigger around their modes, suggesting lower errors for households with higher consumption. This finding is consistent with the expectation that richer households may have better cognitive abilities. Notwithstanding these patterns, the shape of error distributions in Panel B appears reasonably stable. Column (2) of Table 5 corroborates this idea by showing the MSE across deciles of true consumption. We conclude that the quantity Y^d/Y^* is centred below one, with variance approximately independent of Y^*. We take this as evidence that logged reports from a diary are affected by classical errors (aside from location), a finding that has important implications for applied research. On the other hand, the normality of the error distribution, an assumption often made in empirical work (see, for example, Brzozowski et al., 2017), is clearly rejected in our data.

Diary data will yield, in general, biased indicators of poverty and inequality in household consumption. For example, if an absolute poverty line is drawn where the true consumption density is rising (to the left of the mode of $f_{Y^*}(y)$ in Figure 4), the count of poor households based on diaries will be biased upward. This result, which is an implication of the "almost classical" properties of diary errors and prior work by Chesher and Schluter (2002) (e.g., see their Table 2), extends to other indexes of inequality such as the Gini coefficient, the generalized entrophy index and the povery gap. The take-away message here is that diary-recall differences revealed by randomizing households to different interview modes may mislead the assessment of which mode works best for measuring inequality.

Means-testing coming from diaries is not bullet proof either, as we show in Panel B of Figure 5. The curve shown for diary here is $P\left(Y^* \leq \eta | Y^d \leq \eta\right)$, which is the probability of classifying correctly households with consumption below η (for example, someone living under $\eta = 8.75$ USD a week). The other curve, for recall, is the probability of classification $P\left(Y^* \leq \eta | Y^r \leq \eta\right)$. Recall data yield uniformely larger probabilities of correct classification. However, it is clear that raw data yield a classification of households only marginally superior to a coin toss when consumption below the third decile is considered ($\eta = 10.38$). This can be seen by noticing that the probabilities in Panel B are always below 60% at the bottom end of the distribution. We conclude that, although we find evidence

[15] Note that this finding is sufficient to reject the assumption of Berkson-type errors $Y^* = Y^r + \epsilon^r$. The hump-shape behavior at the lower end of the recall curves suggests that the likelihood of reporting zero expenditure is higher as household consumption shrinks to lowest percentiles. As true consumption is bounded at zero, the error distribution cannot have the same support across households.

that recall data are better suited for poverty measurement, a large amount of measurement error remains.

6.2 Consumption Measurements and the Frequency of Purchases

The most likely explanation for the poor performance of diary data is the frequency of purchases, as we show in Figure 8. Panel A here shows diary errors for spending on items which are purchased by households more than once, on average, during the survey week. We expect consumption closer to spending for these items. Panel B of the same figure replicates the analysis for all remaining food items.[16] Diary error distributions in Panel A became less dispersed as true consumption increases, and have a mode closer to one (no error) compared to Figure 6. This implies less dispersion around the mean, which is also one because of Assumption 3. This can be seen from column (5) of Table 5, where presented is the MSE for these distributions, and from columns (4) to (9) of Table 4. Diary errors here show an inverse relationship with total consumption, with the MSE at the top decile being about one-fourth of that in the bottom decile. The densities for items purchased with lower frequency, in Panel B of Figure 8, are more dispersed and have little variation across consumption deciles. Column (8) of Table 5 summarizes these patterns using the MSE.

We conclude that diary errors arise because of two components. The first component depends on the frequency of purchases, which affects diary reports across all values of latent consumption. Once these effects are controlled for, the remaining errors are significantly smaller and with a gradient reflecting true consumption. Differential cognitive abilities of respondents across consumption percentiles can explain this second component. We note, however, that households can sort differently across quantiles of the true distributions for items in Panel A and Panel B of Figure 8. Thus, an alternative interpretation consistent with our findings is that diary respondents tend to underreport those items purchased more frequently but consumed in a smaller amount.

Recall errors tend to be lower for items that are purchased more often. The impact of the frequency of purchases is neverthess less important than in the case of diaries, as shown in the two panels of Figure 9. The distributions here imply that larger consumption yields smaller errors and less overreporting, and more so for high frequency items. Recall error averages are in columns (4) and (7) in Table 5. Values of the MSE are in columns (6) and (9).

7 Implications for the Design of Household Surveys

Is there an optimal assignment of households to diary and recall interviews? The error distributions described above show that neither of these collection modes yields data that are of uniformly better quality. We therefore combine diary and recall responses to determine an assignment rule yielding the minimum loss of accuracy in the distribution of observed consumption.[17]

We start by considering assignments that depend on the true latent consumption. The rationale for using this rule responds to a very practical question: if one knew the true household consumption Y^* , which collection mode would yield its best measurement? Consider a setting where, at each value y^* of Y^*, households are assigned a diary with probability $p(y^*) \in [0, 1]$. The distribution observed in

[16] The classification of items reflects the frequency of purchases in the 2007 and 2012 diaries from the IHSES, and is described in the Appendix. The definition of true consumption in Figure 8, and in Figure 9 below, is for the items considered in each panel (e.g., consumption of items which are purchased more frequently).

[17] Recent research has 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). In all cases, the optimal design choice is determined by minimizing a suitably defined loss function, as we do below.

the data arising from this design is:

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

and the share of survey participants filling out a diary is:

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

A recall survey has $p = 0$, so that all households respond to recall questions independently of their true consumption ($p(y^*) = 0$). Similarly, in a diary survey we have $p = 1$.

We are interested in the effects of the interview mode on functionals of the distribution of observed consumption, $\nu(F_Y(y))$. Possible choices for ν are quantiles, share below the poverty line, and the Gini coefficient. The difference:

$$\nu(F_Y(y)) - \nu(F_{Y^*}(y^*))\,,$$

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, $\nu(F_Y(y))$. We compute (6) from the distributions estimated above, and find the configuration of weights $p(y^*)$ that minimizes the distance between the two statistics at any given level of p (the algorithm used to obtain the solution is described in the Appendix).

A diary survey yields consumption distributions with much larger errors than a recall survey. This can be seen from Panel A of Figure 10, where the statistic ν represents the Kullback-Leibler distance (KLD) from the true distribution $F_{Y^*}(y^*)$. The continuous line in the graph shows the KLD resulting from the allocation in (6) which is optimal subject to the constraint (7). 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 the value $p = 0.42$. If the interview mode could be determined using the true household consumption, the optimal strategy for obtaining the empirical distributions closest to the true distribution would be an approximately equal mix of diary and recall interviews. In particular, we show in the Appendix Figure A.1 that weights $p(y^*)$ at the minimum value of the continuous line are such that households with large values of Y^* should always be assigned to diaries. Second, when $p = 1$ and all households are assigned a diary, the KLD is larger than in the case of a recall survey. This implies that, in our data, the consumption distribution obtained from a fully-fledged recall survey has better properties that distribution obtained from diaries alone.

Errors in consumption distributions obtained from diaries translate into worse measurements of poverty and inequality. This can be seen from the remaining panels of Figure 10, where the statistics $\nu()$ considered are the Gini coefficient (Panel B), the standard deviation of logged consumption (Panel C), and the share of households in the bottom decile (Panel D). The continuous lines in these panels can be interpreted as above. Diary surveys 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 more resources allocated to diaries would improve inequality measurement only until 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 implies very similar conclusions for poverty measurement.

A recall survey is the best option when little information on true household consumption is available. This can be seen from the dashed lines in Figure 10, which are obtained by setting $p(y^*) = p$ in (6) and allocating to diaries a random share p of households. These lines refer to the worst-case scenario where no information on Y^* is employed. 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 true consumption. Proxies of this variable, like income, are often available and used explicitly in the survey design. For example, PSUs are defined to create homogenous groups of households in terms of their income and access to markets. The case of imperfect measurements of Y^* most likely falls in between the dashed and continuous lines in the figure.

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

8 Conclusion

Using a large-scale randomization in Iraq, we have found little empirical support for the idea that diaries yield data of better quality for assessing household welfare. While diaries provide a more reliable measurement of consumption averages, we have shown that the cognitive errors arising from the process of recalling consumption lead to overstate average consumption. However, when inequality and poverty measurement is of interest, the benefits of diaries are far less clear-cut. Diary measurements, despite being correct on average, have large variance.[18] We have argued that this is not the consequence of measurement errors, but mostly the reflection of heterogeneous frequency of purchases across households. We have found that recall modules provide a more reliable mode of collection for inequality and poverty indexes. 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 parts of the distribution. Most national budget surveys use a mix of diary and recall methods, as discussed in many studies (see, for example, Silberstein and Scott, 2011). Nevertheless, our calculations have showed that a fully-fledged recall survey can yield consumption distributions which 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. More research is however needed to assess what makes a good recall module, given that the length of recall list will affect the propensity of respondents to engage in the survey. The decision between using diaries of recall modules is not confined to developing economies. For example, the problems associated with the CEX in the United States have intensified the call for a redesign of the survey, which is underway following the recommendations of a specially appointed NAS 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 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 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 was 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, like in Canada and the United States. The methodology we have presented could

[18] When the between-group component of inequality is of interest, however, diaries still provide the most reliable measurements.

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.

Attanasio, Orazio P. and Luigi Pistaferri, "Consumption Inequality over the Last Half Century: Some Evidence Using the New PSID Consumption Measure," *American Economic Review*, May 2014, *104* (5), 122–26.

_ and _ , "Consumption Inequality," *Journal of Economic Perspectives*, May 2016, *30* (2), 3–28.

_ and Steven J. Davis, "Relative Wage Movements and the Distribution of Consumption," *Journal of Political Economy*, 1996, *104* (6), 1227–1262.

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.

Blundell, Richard, Luigi Pistaferri, and Ian Preston, "Consumption Inequality and Partial Insurance," *American Economic Review*, December 2008, *98* (5), 1887–1921.

_ , _ , and Itay Saporta-Eksten, "Consumption Inequality and Family Labor Supply," *American Economic Review*, February 2016, *106* (2), 387–435.

Bound, John, Charles Brown, and Nancy Mathiowetz, "Measurement Error in Survey Data," in James J. Heckman and Edward Leamer, eds., *Handbook of Econometrics*, Vol. 5, Elsevier, 2001, pp. 3705 – 3843.

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.

Chen, Xiaohong, Han Hong, and Denis Nekipelov, "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, December 2011, *49* (4), 901–37.

Chesher, Andrew, "The effect of measurement error," *Biometrika*, 1991, *78* (3), 451–462.

_ **and Christian Schluter**, "Welfare Measurement and Measurement Error," *The Review of Economic Studies*, 2002, *69* (2), 357–378.

Coibion, Olivier, Yuriy Gorodnichenko, and Dmitri Koustas, "Consumption Inequality and the Frequency of Purchases," Working Paper 23357, National Bureau of Economic Research April 2017.

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

Crain, Bradford R., "Estimation of Distributions Using Orthogonal Expansions," *The Annals of Statistics*, 1974, *2* (3), 454–463.

_ , "An Information Theoretic Approach to Approximating a Probability Distribution," *SIAM Journal on Applied Mathematics*, 1977, *32* (2), 339–346.

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 Salman Zaidi, "Guidelines for Constructing Consumption Aggregates for Welfare Analysis," Technical Report, 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.

Geman, Stuart and Chii-Ruey Hwang, "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 1982, *10* (2), 401–414.

Gibson, John, Kathleen Beegle, Joachim De Weerdt, and Jed Friedman, "What does Variation in Survey Design Reveal about the Nature of Measurement Errors in Household Consumption?," *Oxford Bulletin of Economics and Statistics*, 2014, *77* (3), 466–474.

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.

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

Kotlarski, Ignacy, "On characterizing the gamma and the normal distribution.," *Pacific J. Math.*, 1967, *20* (1), 69–76.

Krueger, Dirk and Fabrizio Perri, "Does Income Inequality Lead to Consumption Inequality? Evidence and Theory1," *The Review of Economic Studies*, 01 2006, *73* (1), 163–193.

Lyberg, Lars and Daniel Kasprzyk, "Data Collection Methods and Measurement Error: An Overview," in "Measurement Errors in Surveys," Wiley-Blackwell, 2011, 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.

Menon, Geeta, "The Effects of Accessibility of Information in Memory on Judgments of Behavioral Frequencies," *Journal of Consumer Research*, 1993, *20* (3), 431–440.

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.

Peterson Zwane, Alix, Jonathan Zinma, Eric Van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, Dean S. Karlan, Richard Hornbeck, Xavier Giné, Esther Duflo, Florencia Devoto, Bruno Crepon, and Abhijit Banerjee, "Being surveyed can change later behavior and related parameter estimates," *Proceedings of the National Academy of Sciences*, 2011, *108* (5), 1821–1826.

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.

Silberstein, Adriana R. and Stuart Scott, "Expenditure Diary Surveys and Their Associated Errors," in "Measurement Errors in Surveys," Wiley-Blackwell, 2011, chapter 16, pp. 303–326.

Table 1: Summary statistics and covariate balance

	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	5.9567	-0.0041	5.7984	-0.0094	6.0638	-0.0006
	[0.6533]	(0.0083)	[0.6752]	(0.0137)	[0.6153]	(0.0105)
Log implicit rent	4.1307	-0.0092	3.5675	-0.0036	4.5649	-0.0137
	[0.8482]	(0.0065)	[0.7526]	(0.0099)	[0.6360]	(0.0086)
Log price index	0.9847	-0.0016	0.9598	-0.0014	1.0015	-0.0017
	[0.1211]	(0.0012)	[0.1134]	(0.0018)	[0.1233]	(0.0016)
Household size	7.0189	0.0406	7.6547	0.0825	6.5887	0.0121
	[3.5678]	(0.0470)	[3.9420]	(0.0847)	[3.2202]	(0.0540)
N. of children	2.3881	0.0037	2.8098	-0.0015	2.1028	0.0072
	[2.0660]	(0.0283)	[2.2667]	(0.0495)	[1.8651]	(0.0336)
N. of adults	4.0281	0.0068	4.1326	0.0362	3.9573	-0.0132
	[2.3175]	(0.0316)	[2.4422]	(0.0529)	[2.2265]	(0.0390)
N. of occupied adults	1.5214	-0.0129	1.5776	-0.0141	1.4834	-0.0121
	[1.0782]	(0.0145)	[1.1864]	(0.0251)	[0.9966]	(0.0175)
Max. education level	3.1920	0.0300	2.7268	0.0593*	3.5068	0.0099
	[1.7517]	(0.0222)	[1.6376]	(0.0338)	[1.7563]	(0.0295)
Panel B. Household head demographics						
Male	0.8999	0.0016	0.9203	-0.0037	0.8862	0.0052
	[0.3001]	(0.0042)	[0.2709]	(0.0061)	[0.3176]	(0.0057)
Age	45.8584	-0.0567	45.0709	-0.2206	46.3915	0.0539
	[13.8499]	(0.1931)	[14.1377]	(0.2996)	[13.6266]	(0.2524)
Education level	1.9906	-0.0103	1.6526	0.0022	2.2193	-0.0186
	[1.8123]	(0.0232)	[1.6477]	(0.0330)	[1.8817]	(0.0319)
Reads and writes	0.7253	0.0017	0.6706	0.0030	0.7623	0.0009
	[0.4467]	(0.0058)	[0.4704]	(0.0094)	[0.4259]	(0.0074)
Employed	0.7454	0.0015	0.7569	-0.0024	0.7377	0.0041
	[0.4359]	(0.0060)	[0.4293]	(0.0090)	[0.4402]	(0.0081)
Number of districts	119	119	119	119	119	119
Number of EAs	2,828	2,827	2,828	2,827	2,828	2,827
Number of households	16,530	8,220	16,530	8,220	16,530	8,220

Note. 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. Diaries are the baseline mode of collection in the Iraq Household and Socio-Economic Expenditure Survey (IHSES). 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

	Baseline	**Treatment**		**Treatment Difference**	
	Diary	Diary	Recall	Diary	Recall
	(1)	(2)	(3)	(4)	(5)
			Panel A. Location and percentiles		
5th percentile	2.5291	2.5131	2.7858	-0.0160	0.2567***
				(0.0192)	*(0.0163)*
25th percentile	2.0622	2.0375	2.3880	-0.0247	0.3258***
				(0.0203)	*(0.0184)*
50th percentile	2.5291	2.5131	2.7858	-0.0160	0.2567***
				(0.0192)	*(0.0163)*
75rh percentile	2.9788	2.9543	3.1787	-0.0245	0.2000***
				(0.0195)	*(0.0171)*
95th percentile	3.6601	3.6187	3.6804	-0.0414	0.0202
				(0.0343)	*(0.0257)*
Mean	2.5184	2.5010	2.7878	-0.0158	0.2706***
				(0.0099)	*(0.0105)*
			Panel B. Dispersion		
Std. deviation	0.7992	0.7944	0.7039	-0.0048	-0.0953***
				(0.0110)	*(0.0156)*
Gini index	0.4104	0.3996	0.3185	-0.0108	-0.0919***
				(0.0071)	*(0.0062)*
Number of districts	119	119	119	119	119
Number of EAs	2,828	2,827	2,827	2,827	2,827
Number of households	16,530	8,220	8,220	8,220	8,220

Note. This table shows statistics from the baseline and the treatment samples. Household is the unit of observation. Diaries are the baseline mode of collection in the Iraq Household and Socio-Economic Expenditure Survey (IHSES). Column (1) reports statistics for the baseline sample, which consists of households excluded from the recall-module experiment. Columns (2) and (3) report the same statistics for the treatment sample, which consists of households randomized to the recall-module experiment. For these households, both diary and recall measurements are available. Differences in columns (4) and (5) of Panel A are coefficients from plain or quantile regressions 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. In these regressions, the outcome in column (4) uses the log of diary reports for both the baseline and the treatment samples. The outcome in column (5) uses the log of diary reports for the baseline sample and log of recall reports for the treatment sample. In Panel B, standard errors for the difference between standard deviations of logs and Gini coefficients in columns (4) and (5) are computed via bootstrap using 1,000 replications. Standard errors (in brackets) are clustered on EAs. * $p < 0:10$, **$p < 0:05$, *** $p < 0:01$.

Table 3: Time of interview effects on demographics and consumption measurements

	Bimester					
	II	**III**	**IV**	**V**	**VI**	**F-test**
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A. Household characteristics					
Age	-0.040	-0.007	-0.033	-0.017	0.008	0.339
	(*0.026*)	(*0.025*)	(*0.025*)	(*0.026*)	(*0.025*)	
Education level	0.036	0.030	0.017	0.048	0.067**	0.294
	(*0.031*)	(*0.030*)	(*0.031*)	(*0.030*)	(*0.030*)	
Employed	-0.010	0.008	0.030	0.004	0.024	0.505
	(*0.023*)	(*0.023*)	(*0.023*)	(*0.023*)	(*0.023*)	
	Panel B. Spending and prices					
Log expenditure	0.204***	0.098***	0.080***	0.109***	0.030	0.000
	(*0.022*)	(*0.021*)	(*0.021*)	(*0.022*)	(*0.022*)	
Log price index	0.026***	0.026***	-0.017***	0.008**	0.007**	0.000
	(*0.004*)	(*0.003*)	(*0.003*)	(*0.003*)	(*0.003*)	
Exp. share on rice	0.180***	0.202***	0.232***	0.025	0.027	0.000
	(*0.022*)	(*0.024*)	(*0.022*)	(*0.021*)	(*0.021*)	
Exp. share on potatoes	-0.146***	-0.162***	-0.245***	-0.135***	-0.059**	0.000
	(*0.027*)	(*0.027*)	(*0.027*)	(*0.029*)	(*0.029*)	
Exp. share on eggs	-0.137***	-0.078***	-0.184***	-0.184***	0.032	0.000
	(*0.025*)	(*0.026*)	(*0.025*)	(*0.026*)	(*0.027*)	
Exp. share on meat	0.012	0.108***	0.222***	0.361***	0.107***	0.000
	(*0.023*)	(*0.023*)	(*0.023*)	(*0.025*)	(*0.024*)	
Exp. share on fish	0.011	-0.047**	-0.088***	-0.088***	-0.039*	0.000
	(*0.022*)	(*0.021*)	(*0.022*)	(*0.021*)	(*0.022*)	

Note. This table shows how household characteristics (Panel A) and diary spending (Panel B) of interviewees change during the year-long survey. Household is the unit of analysis. Diaries are the baseline mode of collection in the Iraq Household and Socio-Economic Expenditure Survey (IHSES). Columns (1) to (5) report coefficients of bimester of interview dummies in regressions of the variables at left that control for survey-design strata (districts). Bimesters are defined to match agricultural seasons: December-January (omitted) and February-March (I) are the lean or off-season; April-May (II) and June-July (III) are the harvest season; August-September (IV) and October-November (VI) are the post-harvest season. Column (6) reports the p-value of the F-test for the joint equality of bimester of interview coefficients. All variables at left are standardized to have zero mean and unit variance, unless logs are used. Standard errors for the coefficients on the bimester of interview dummies (in brackets) are clustered on EAs. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Survey measurements and true consumption

				Frequency of Purchases					
	All Food Items			**Higher**			**Lower**		
	Diary	Recall	True	Diary	Recall	True	Diary	Recall	True
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1st decile	4.5821	7.2327	7.2632	1.0250	1.6975	1.4516	2.8114	4.7312	5.1699
2nd decile	6.5176	9.7270	8.9081	1.4815	2.1413	1.8013	4.5821	6.9488	6.7639
3rd decile	8.2875	11.8830	10.3819	1.8687	2.5132	2.1042	6.2618	8.9069	8.2744
4th decile	10.0437	14.0592	11.7948	2.2467	2.8797	2.3951	8.0263	10.8810	9.7410
5th decile	12.0749	16.2391	13.3999	2.6369	3.2473	2.7029	10.0437	12.9773	11.3579
6th decile	14.4011	18.9078	15.0945	3.0702	3.6617	3.0502	12.4679	15.4773	13.2432
7th decile	17.3135	22.0151	17.2951	3.5748	4.1623	3.4421	15.4773	18.4590	15.4415
8th decile	21.4925	26.0468	20.1566	4.2635	4.8463	3.9863	19.8385	22.7317	18.7094
9th decile	28.9046	33.1199	24.9328	5.4212	5.9205	4.9042	27.7701	29.8454	24.2443
Mean	15.1609	18.8219	15.1539	2.9658	3.5618	2.9645	13.5104	15.6988	13.4616
Std. deviation	12.0195	11.5340	7.9407	1.8272	1.7798	1.4428	12.7706	11.3004	8.6616
Gini index	0.3792	0.3123	0.2644	0.3253	0.2630	0.2556	0.4377	0.3636	0.3212
Number of districts	119	119	119	119	119	119	119	119	119
Number of EAs	2,828	2,827	2,827	2,828	2,827	2,827	2,828	2,827	2,827
Number of households	24,750	8,220	8,220	24,750	8,220	8,220	24,750	8,220	8,220

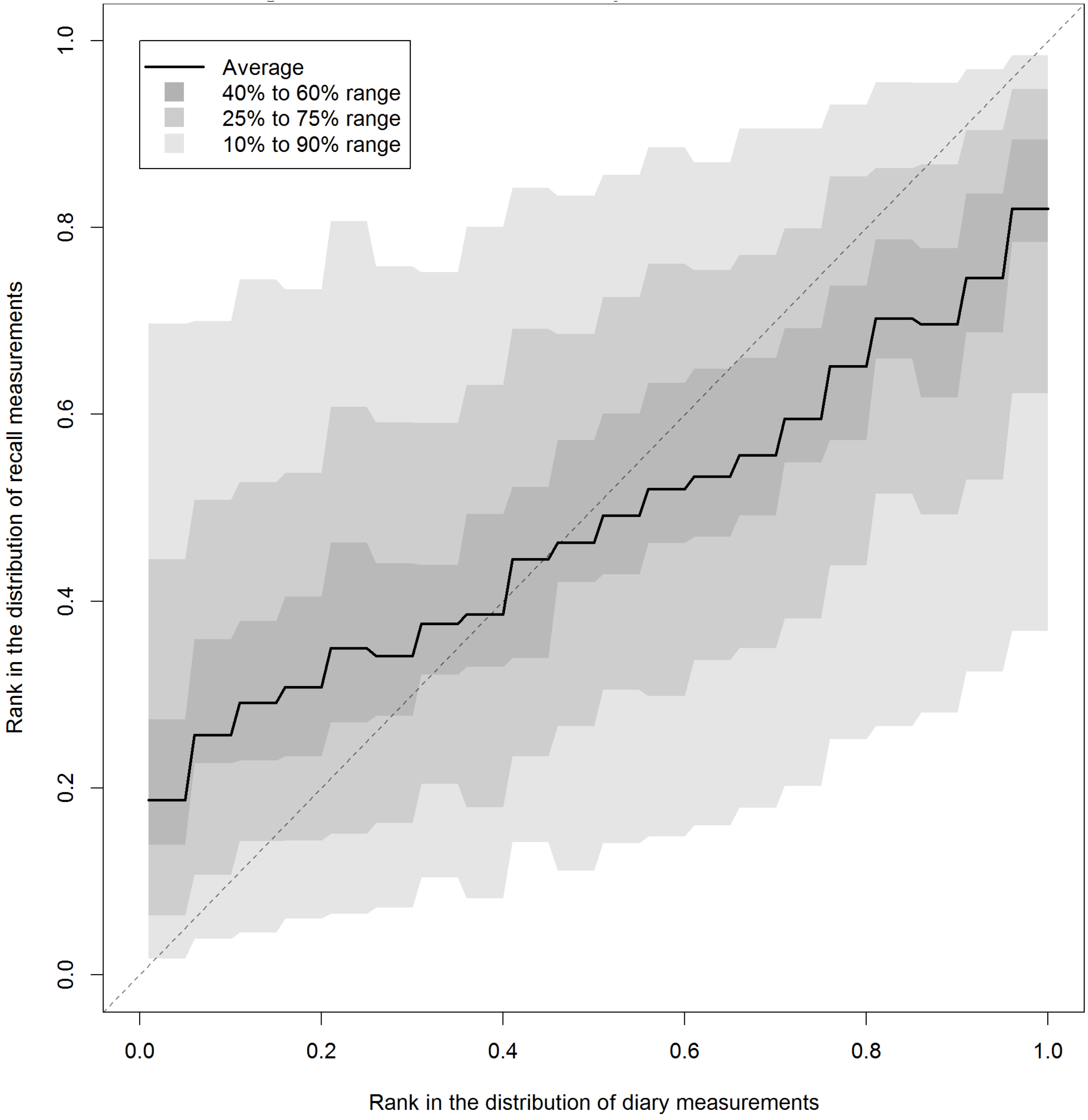
Note. This table compares statistics of food reports in the baseline and the treatment samples to statistics of the true food consumption distribution. Household is the unit of observation. Diaries are the baseline mode of collection in the Iraq Household and Socio-Economic Expenditure Survey (IHSES). Columns (1), (4) and (7) use diaries for all households. Columns (2), (5) and (8) use recall data in the treatment sample, which consists of households randomized to the recall-module experiment. Columns (3), (6) and (9) report the statistics estimated from our model. Columns (1) to (3) consider figures for total consumption obtained by aggregating all food items. Columns (4) to (9) consider totals obtained by aggregating items that are purchased frequently (i.e., bought more than once during the diary week) and infrequently (less than once during the diary week).

Table 5: Summary statistics for the error distributions

				Frequency of Purchases					
	All Food Items			**Higher**			**Lower**		
	Mean	**MSE**		**Mean**	**MSE**		**Mean**	**MSE**	
	Recall	Diary	Recall	Recall	Diary	Recall	Recall	Diary	Recall
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1st decile	1.411	0.335	0.708	1.830	0.326	1.282	1.402	0.328	0.738
2nd decile	1.342	0.322	0.530	1.563	0.238	0.779	1.386	0.321	0.654
3rd decile	1.301	0.314	0.436	1.412	0.192	0.550	1.358	0.315	0.590
4th decile	1.272	0.308	0.375	1.309	0.164	0.418	1.327	0.309	0.539
5th decile	1.248	0.302	0.327	1.229	0.144	0.330	1.293	0.305	0.491
6th decile	1.229	0.296	0.291	1.162	0.127	0.266	1.255	0.301	0.446
7th decile	1.211	0.289	0.256	1.107	0.114	0.220	1.211	0.299	0.402
8th decile	1.194	0.282	0.226	1.051	0.102	0.180	1.150	0.299	0.351
9th decile	1.176	0.270	0.193	0.990	0.088	0.144	1.055	0.303	0.293
Number of districts	119	119	119	119	119	119	119	119	119
Number of EAs	2,828	2,827	2,827	2,828	2,827	2,827	2,828	2,827	2,827
Number of households	8,220	8,220	8,220	8,220	8,220	8,220	8,220	8,220	8,220

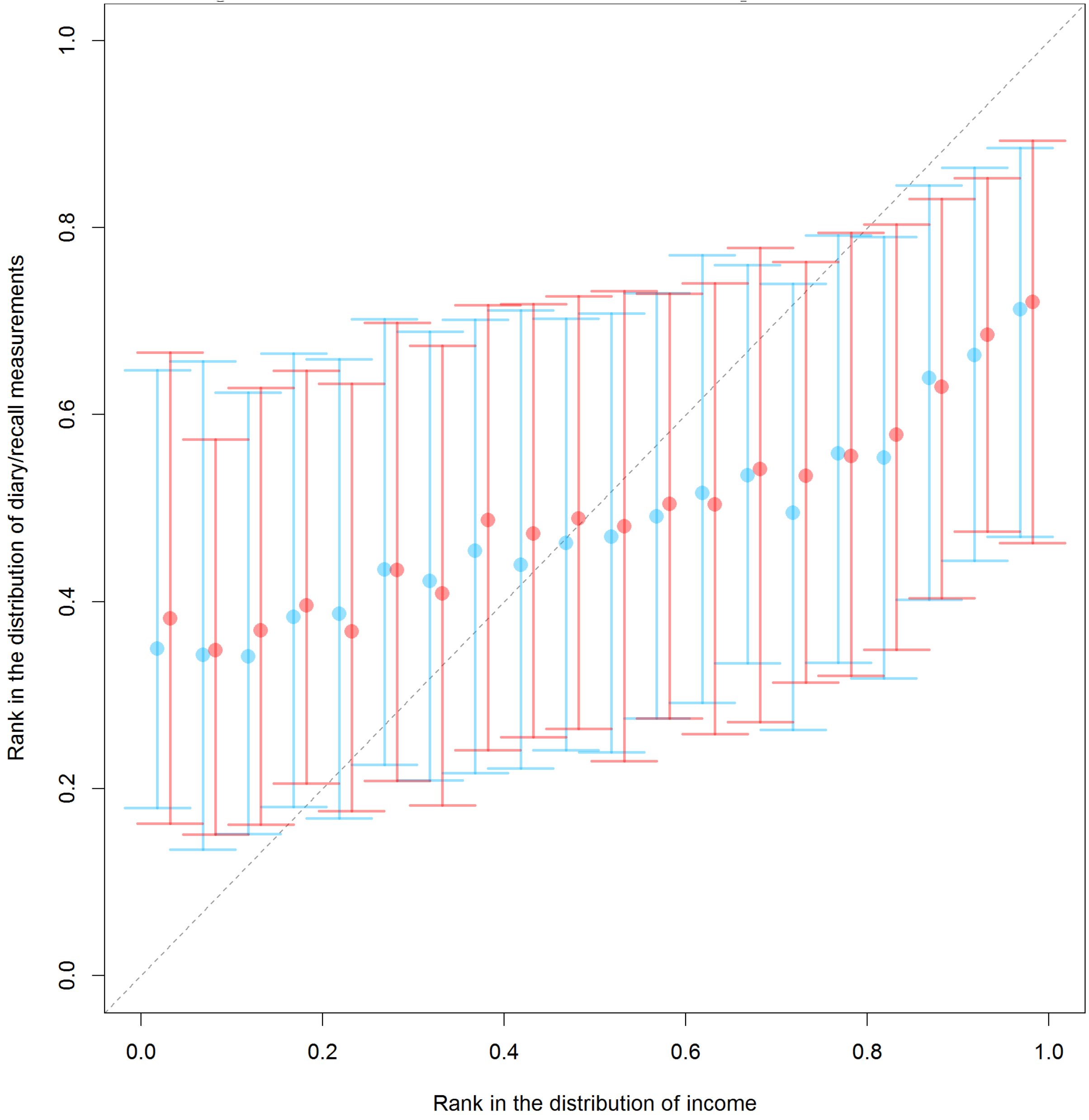
Note. This table shows statistics of the conditional distributions of Y^r/Y^* (recall) and Y^d/Y^* (diary) at selected percentiles of Y^* (true consumption, at left). The conditional means of Y^r/Y^* are in columns (1), (4) and (7). All conditional means of Y^d/Y^* are one by construction. The remaining columns report mean squared errors of the conditional distributions. Columns (1) to (3) consider figures for total consumption obtained by aggregating all food items. Columns (4) to (9) consider totals obtained by aggregating items that are purchased frequently (i.e., bought more than once during the diary week) and infrequently (less than once during the diary week).

Figure 1: Household rank in diary and recall distributions



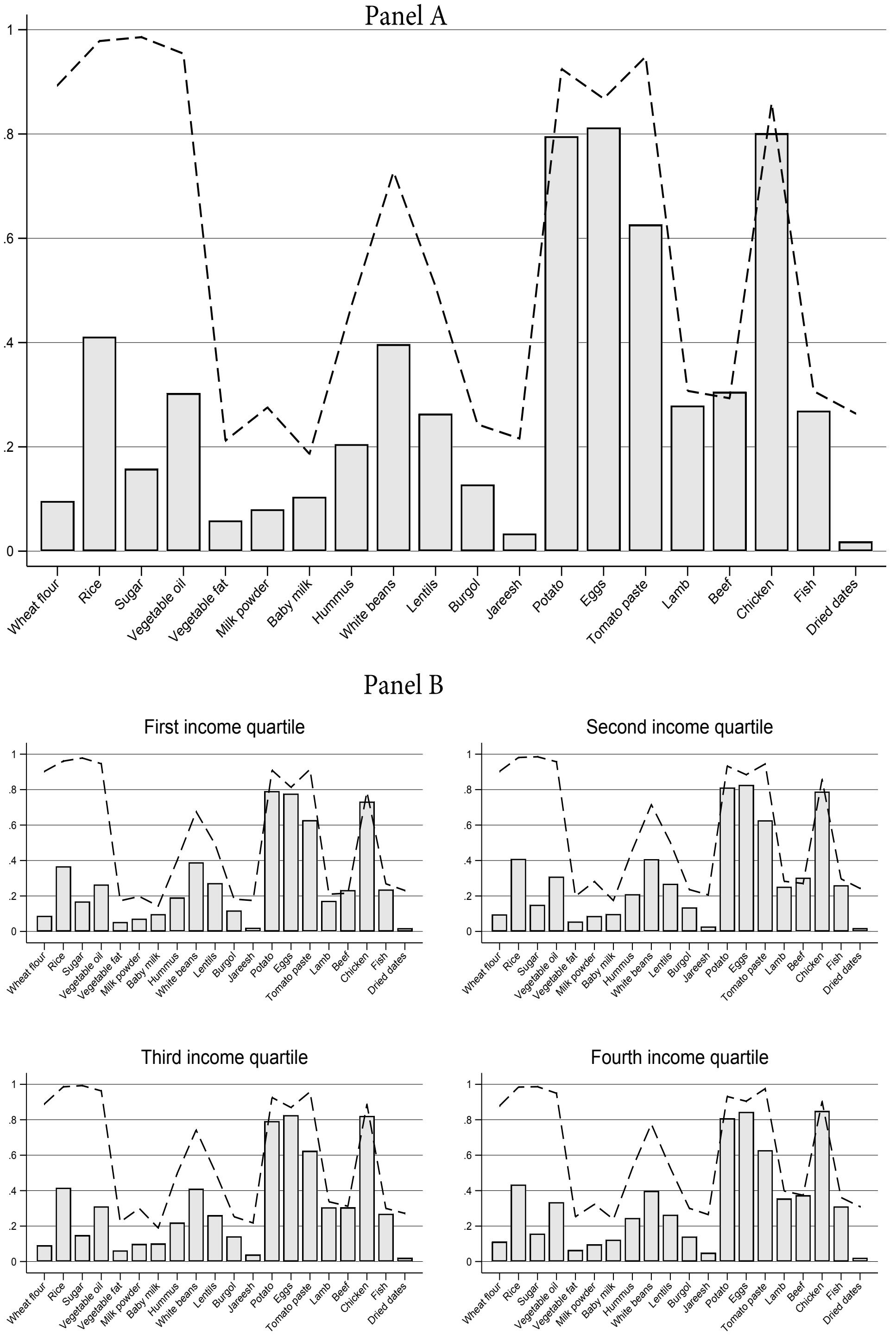
Note. This figure compares diary and recall reports from the treatment sample, which consists of households randomized to the recall-module experiment. Households are ranked in diary and recall distributions depending on their reports. For households falling in the p-th percentile of the diary distribution on the horizontal axis, the vertical axis shows summary statistics for the percentile of the same households in the recall distribution. Rank invariance (same rank in both distributions) corresponds to the dashed line in the figure (the 45-degree line). The continuous line represents the average percentile in the recall distribution for households falling in the p-th percentile of the diary distribution. The darker area represents the 40% to 60% range in the recall distribution for households falling in the p-th percentile of the diary distribution. The remaining colors report the 25% to 75% range and the 10% to 90% range.

Figure 2: Household rank in income and consumption distributions



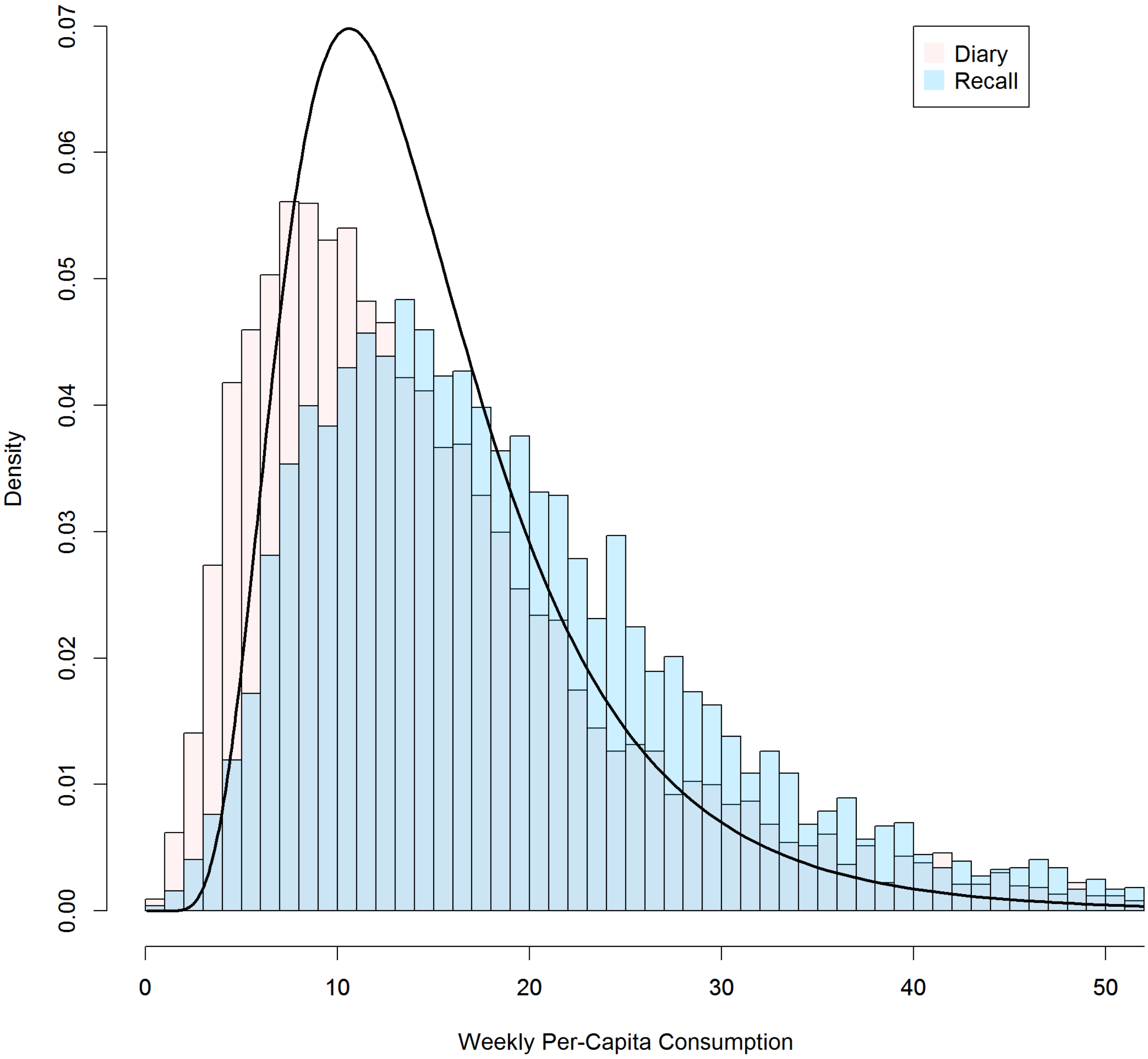
Note. This figure shows the correlation of diary and recall reports with household income using the treatment sample, which consists of households randomized to the recall-module experiment. Households are ranked in diary and recall distributions depending on their reports. For households falling in the p-th percentile of the income distribution on the horizontal axis, the vertical axis shows summary statistics for the percentile of the same households in the diary and recall distributions. Rank invariance (same rank in income and consumption distributions) corresponds to the dashed line in the figure (the 45-degree line). The continuous lines are the average percentile in the diary or recall distributions for households falling in the p-th percentile of the income distribution. The figure also reports the 25% to 75% consumption range plots for households falling in the p-th percentile of the income distribution.

Figure 3: Consumption and spending across food items



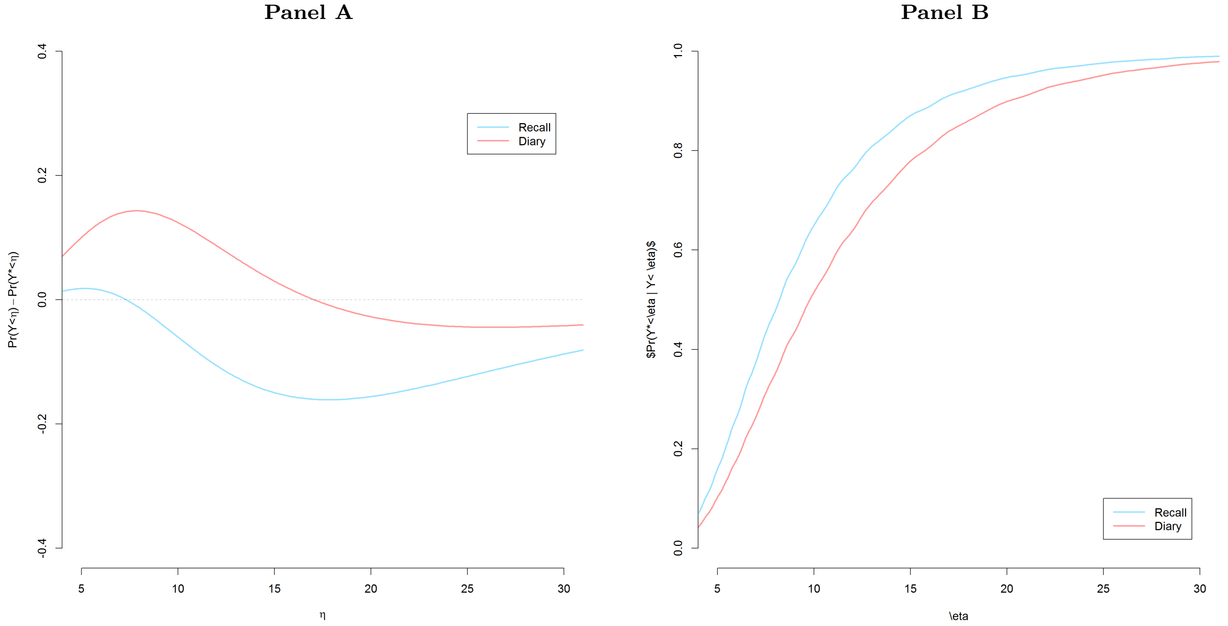
Note. Panel A shows the share of households with positive spending in the diary week (in bars) and the share of households self-reporting positive consumption in the recall module (the dashed line). The 20 consumption groups used in the recall module are reported on the horizontal axis. Panel B shows the same figures after a stratification by quartile of the income distribution.

Figure 4: Survey measurements and true consumption



Note. This figure shows the empirical densities of food measurements (histograms) and the estimated density of true consumption (continuous line). Household is the unit of observation. Diaries are the baseline mode of collection in the Iraq Household and Socio-Economic Expenditure Survey (IHSES). The histogram for diaries is computed using all households in the survey. The histogram for recall uses the treatment sample, which consists of households randomized to the recall-module experiment. The continuous line is estimated using the model described in Section 5.

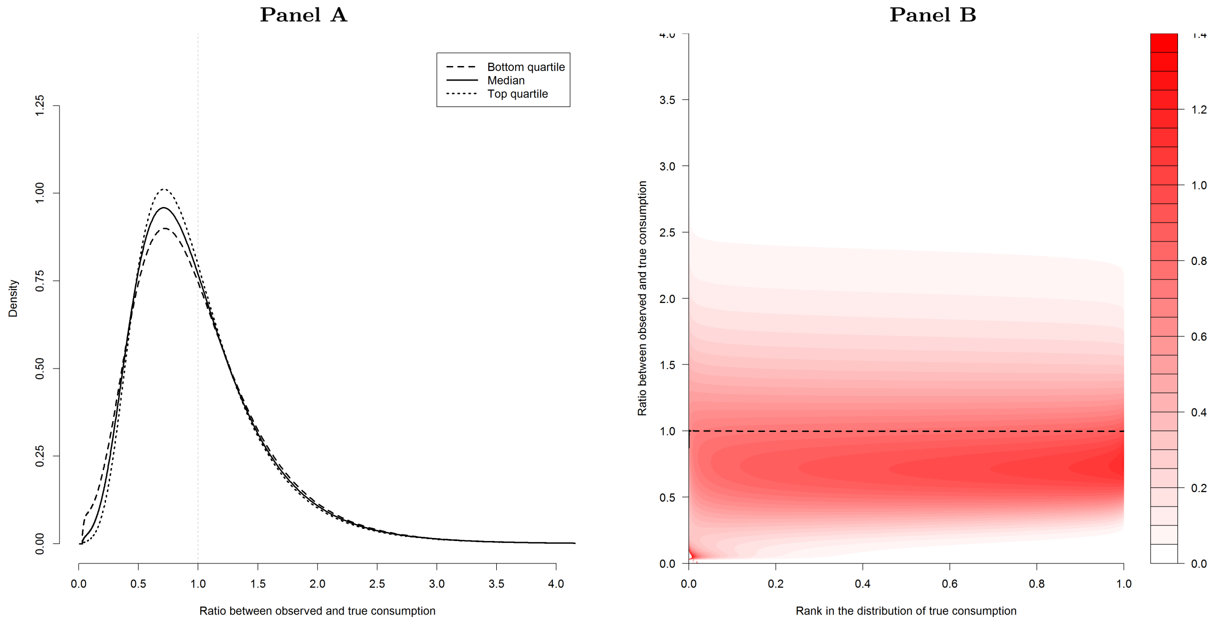
Figure 5: Poverty mismeasurement and the misclassification of households


Panel A

Panel B

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 true distributions, $P\left(Y^* \leq \eta\right)$. 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.

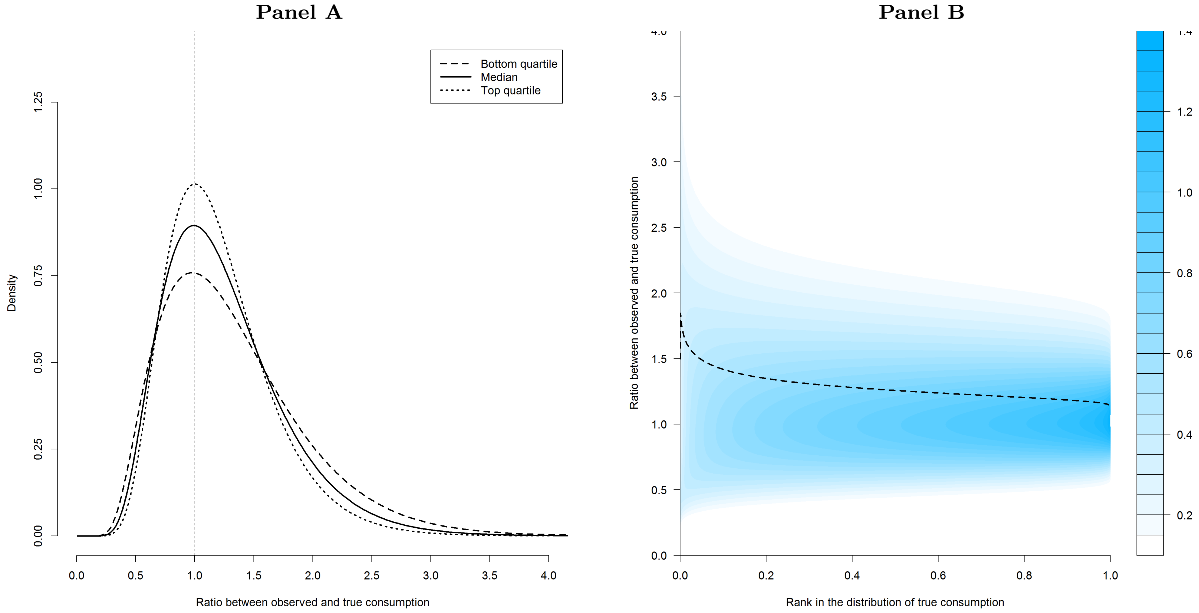
Figure 6: Survey effects on measured consumption using diaries


Panel A

Panel B

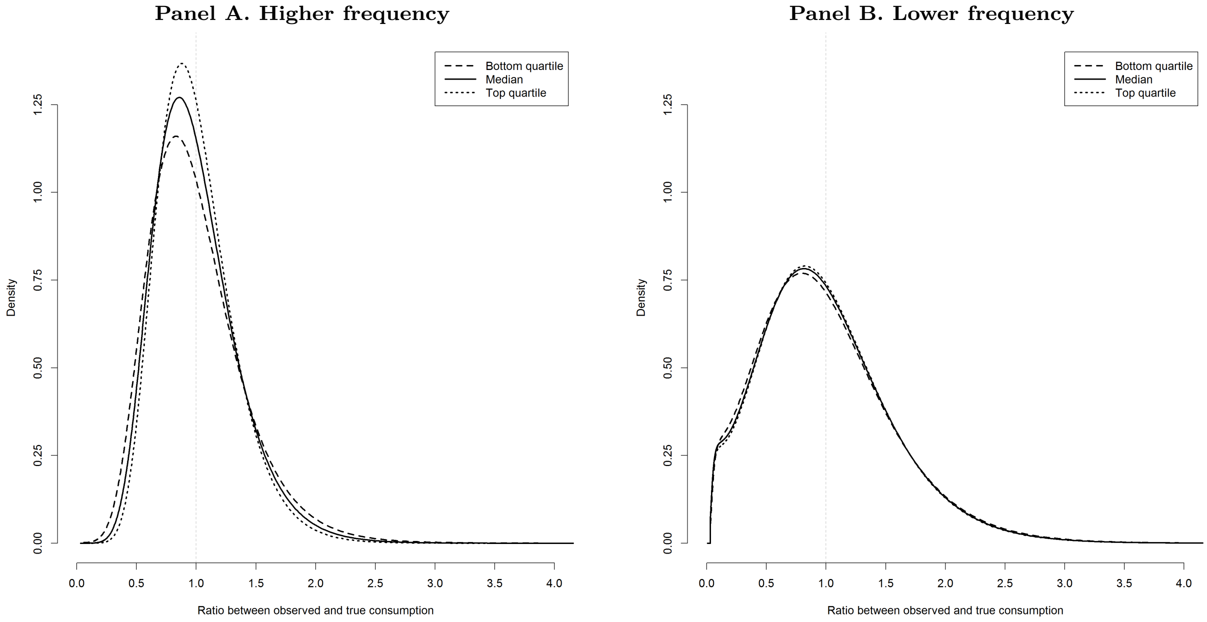
Note. Panel A shows error distributions for a hypothetical household at three selected percentiles of true consumption, Y^*. The quantity Y^d/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from diaries, Y^d, and her true consumption. No diary error implies 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 (i.e., no error on average).

Figure 7: Survey effects on measured consumption using recall questions



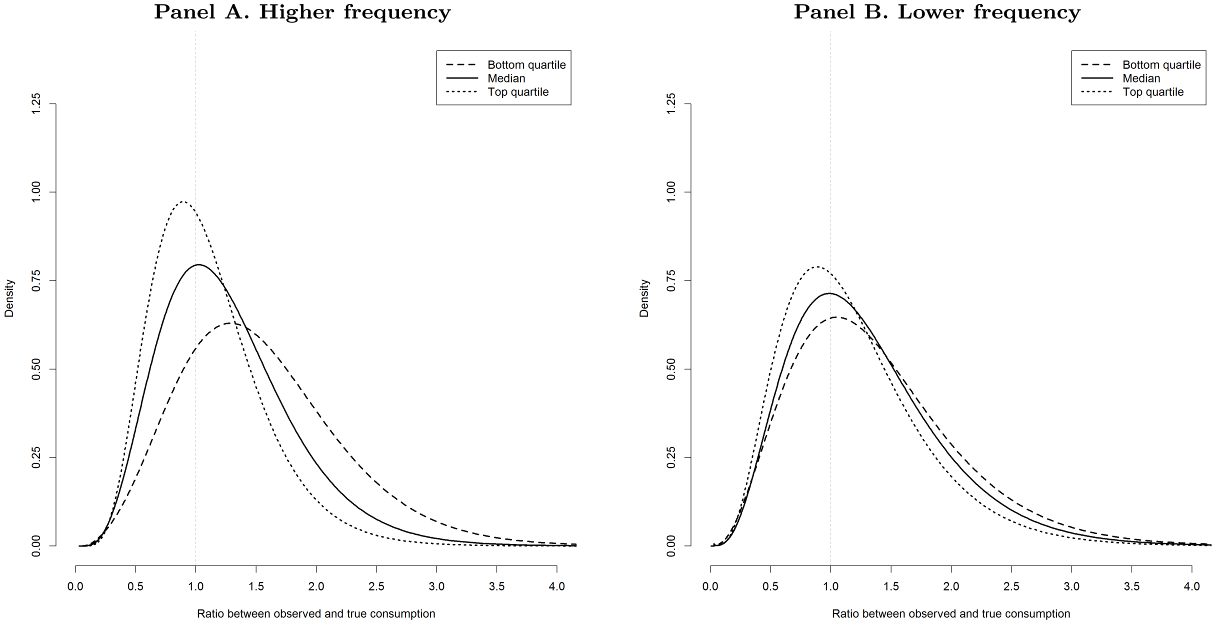
Note. Panel A shows error distributions for a hypothetical household at three selected percentiles of true consumption, Y^*. The quantity Y^r/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from the recall module, Y^r, and her true consumption. No recall 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.

Figure 8: Survey effects from diaries and frequency of purchases



Note. Panel A shows error distributions for a hypothetical household at selected percentiles of true consumption, Y^*, for food items purchased frequently in the diary week. Panel B shows error distributions considering food items purchased less frequently in the diary week. The quantity Y^d/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from diaries, Y^d, and her true consumption. No diary error implies a value of one for this ratio.

Figure 9: Survey effects from recall questions and frequency of purchases



Note. Panel A shows error distributions for a hypothetical household at selected percentiles of true consumption, Y^*, for food items purchased frequently. Panel B shows error distributions considering food items purchased less frequently. The quantity Y^r/Y^* on the horizontal axis denotes the relationship between the household's survey measurement from recall questions, Y^r, and her true consumption. No recall error implies a value of one for this ratio.

Appendix

Optimal allocation of diary and recall interviews

We describe here how to determine weights $p(y^*)$ in Section 7 which minimize the Kullback-Leibler distance (KLD) between the observed and true consumption distributions.

We start by writing $p(y^*)$ as:

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

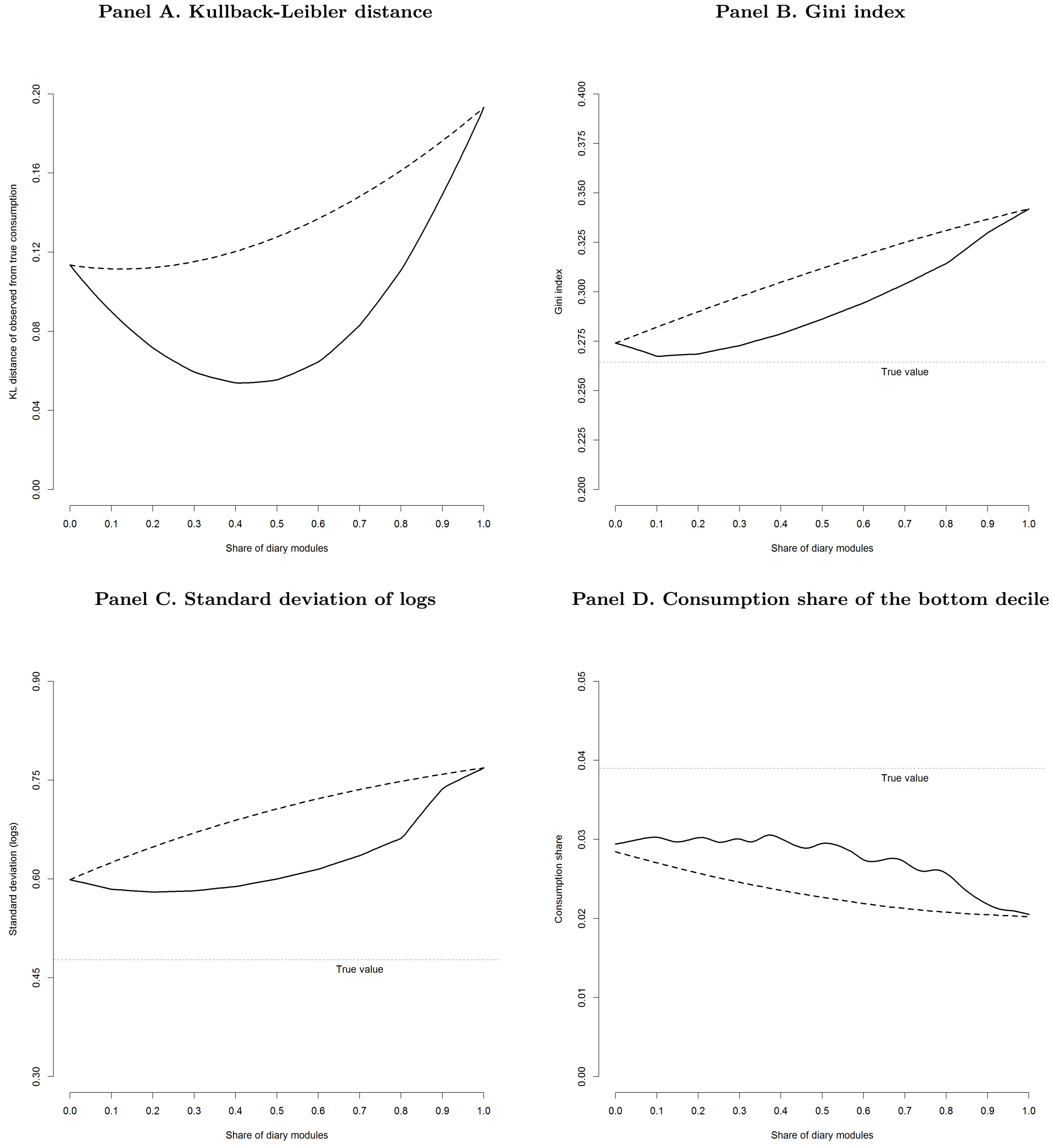
where $y^*_{(0)}, y^*_{(1)}, \ldots, y^*_{(10)}$ are the deciles of the distribution of Y^* and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{10})'$. The last expression defines a flexible piecewise-constant function for $p(y^*)$, with α_i representing the probability of being assigned a diary module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

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

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

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

Figure 10: Optimal allocation of respondents to interview modes



Panel A. Kullback-Leibler distance

Panel B. Gini index

Panel C. Standard deviation of logs

Panel D. Consumption share of the bottom decile

Note. The horizontal axis in the figure shows the share p of diary interviews in the design. The value $p = 0$ corresponds to a recall survey (no diaries). The value $p = 1$ corresponds to a fully-fledged diary survey (only diaries). The continuous lines are for empirical distributions from allocating to diaries a share p of households using their true consumption Y^* – see equations (6) and (7) in Section 7. The dashed lines are obtained by considering the empirical distributions from allocating to diaries a share p of households selected at random. Panel A shows the Kullback-Leibler distance between the true distribution of consumption and the empirical distributions resulting under these allocations. Panel B shows the difference between the true Gini index and the Gini index resulting under these allocations. The value of the true index is also reported. Lines in Panel C (standard deviation of longs) and Panel D (consumption share of the bottom decile) should be interpreted as in Panel B.

and:

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

This is a constrained optimization problem over $\boldsymbol{\alpha}$. The results are shown in Figure A.1. Panel B here 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 true consumption distribution and recall modules with probability one to the remaining households. Consequently we 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 true consumption density is plotted alongside the observed densities obtained by assigning only diaries, only recall modules or the optimal survey design.

In order to explore the optimal survey design under alternative assigned proportions of diary modules (p) we also consider:

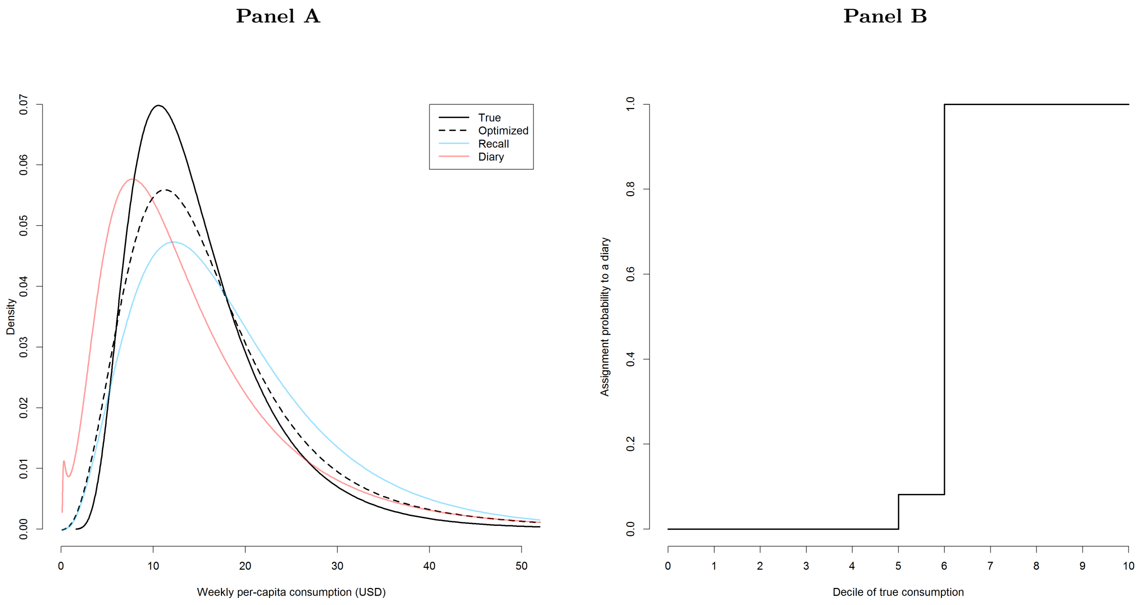
$$\begin{aligned} \hat{\boldsymbol{\alpha}}_p &= \underset{\boldsymbol{\alpha}}{\text{argmax}} \; \nu(F_Y(y;\boldsymbol{\alpha})) \\ &\text{s.t.} \;\; 0 \le \alpha_i \le 1, \;\text{ for } i = 1, \ldots, 10 \\ &\text{s.t.} \;\; \int_0^\infty p(y^*;\boldsymbol{\alpha})dF_{Y^*}(y^*) = p \end{aligned}$$

This allows to estimate the optimal survey design which allocates a share p of the households to diaries. Functionals of the observed distributions under alternative choices of p, i.e.

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

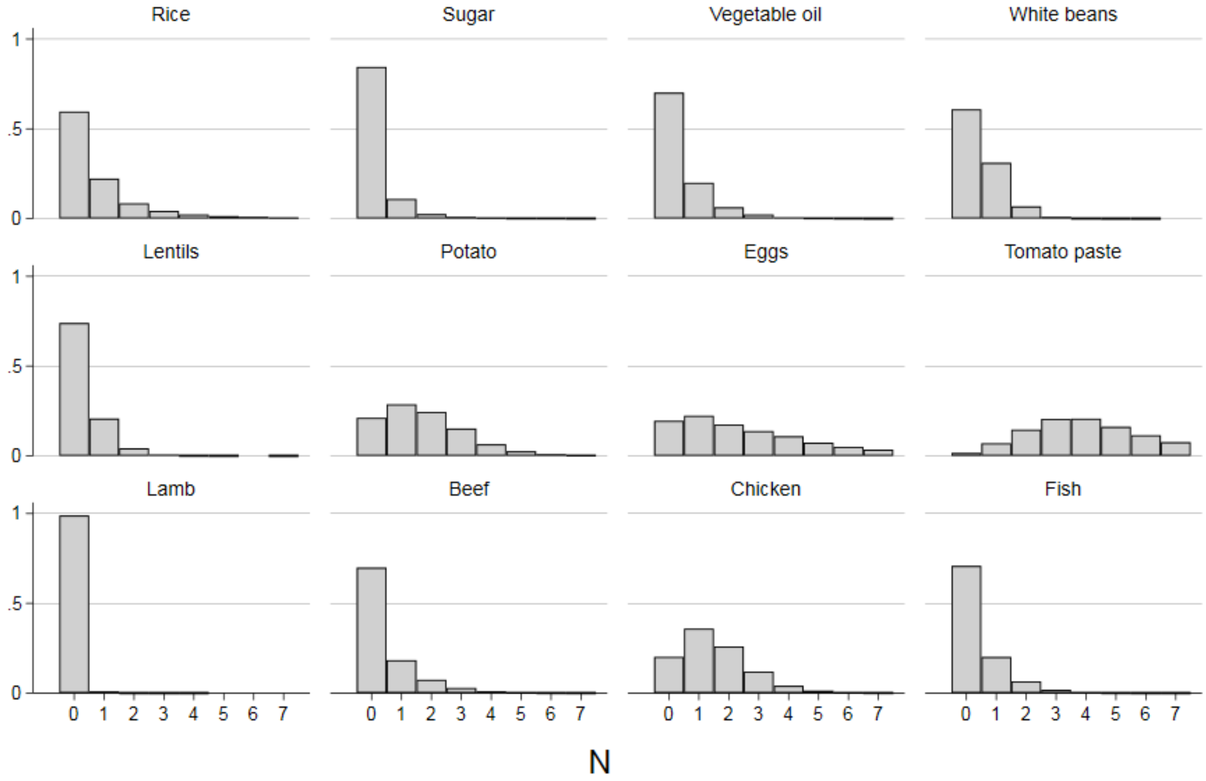
are reported in Figure 10.

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 5 to obtain the true distribution (True); (d) the allocation in Section 7 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.

Figure A.2: Frequency of purchases across items



Note. This figure presents the distributions of the number of times each item was purchased in the diary week.