November 2011

Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania

Forthcoming in Journal of Development Economics

Kathleen Beegle, World Bank
Joachim De Weerdt, EDI Tanzania
Jed Friedman, World Bank
John Gibson, University of Waikato

Abstract

Consumption expenditure has long been the preferred measure of household living standards. However, accurate measurement is a challenge and household expenditure surveys vary widely across many dimensions, including the level of reporting, the length of the reference period, and the degree of commodity detail. These variations occur both across countries and also over time within countries. There is little current understanding of the implications of such changes for spatially and temporally consistent measurement of household consumption and poverty. A field experiment in Tanzania tests eight alternative methods to measure household consumption on a sample of 4,000 households. There are significant differences between consumption reported by the benchmark personal diary and other diary and recall formats. Under-reporting is particularly relevant in illiterate households and for urban respondents completing household diaries; recall modules measure lower consumption than a personal diary, with larger gaps among poorer households and households with more adult members. Variations in reporting accuracy by household characteristics are also discussed and differences in measured poverty as a result of survey design are explored. The study concludes with recommendations for methods of survey based consumption measurement in low-income countries.

Keywords: consumption, expenditure, survey design
JEL: C83, D12

* This work was supported by the World Bank Research Committee. We would like to thank two anonymous referees, seminar participants at IFPRI, the ASSA annual meetings, and the Survey Design and Measurement in Development Economics conference for very useful comments. All views are those of the authors and do not reflect the views of the World Bank or its member countries.

1. Introduction

Household consumption is typically the core concept at the center of any attempt to measure living standards, inequality and poverty in the developing world.[1] Measures of consumption are derived almost exclusively through survey, however a review of survey practice will quickly reveal extensive variation over several dimensions, such as the method of data capture (diary versus recall), the level of respondent (individual versus household), the reference period for which consumption is reported (anywhere from 3 days to one year) and the degree of commodity detail (from less than 20 items to over 400 items). Variations occur both across countries and over time as statistical offices alter survey designs, with little understanding of the implications of such changes for accurate and consistent measurement of consumption. Such variation hampers both cross-country studies of poverty and inequality measures as well as the measurement of welfare trends within a country (Lanjouw and Lanjouw, 2001).[2] For example, the national living standards survey in South Africa found a decline in real food expenditure from 2000 to 2005 even though the same data sources indicated increases in overall consumption and declines in self-reported hunger/food insecurity. This contradictory finding is largely attributed to a change in survey design from recall to diary for food consumption measurement in 2005 (Yu, 2008).

Changes in survey design, if they matter, need to be considered in analysis of trends in consumption or poverty over time. By extension, analysis that pools surveys of different types to study economic growth and inequality should take these differences seriously. Consider, for example, the widely used UNU/WIDER World Income Inequality Database (WIID) on income distribution and inequality. Concerns about consistency in such cross-country databases focus on whether the source surveys are income or consumption surveys (such as Atkinson and Brandolini, 2001, and Pinkovskiy and Sala-i-Martin, 2009), and the comprehensiveness of the consumption measure (Chen and Ravallion, 2010), but not on the implications of variation in consumption survey design. Although we do not attempt a complete inventory of the consumption surveys in this database, for the data of six African countries (chosen on the basis of our past experience or access to survey materials: Côte d`Ivoire, Ghana, Malawi, South Africa, Tanzania, and Zambia)

[1] Conceptual and practical reasons favoring consumption expenditures over income as a welfare measure are discussed in Deaton (1997). Empirically, statistics offices in a majority of the developing countries, with metadata in the United Nations Handbook of Poverty Statistics (UN 2005), use either consumption expenditures solely or in combination with income as their welfare measure. The only region with a high reliance on income surveys is Latin America, although even there increased use is being made of expenditure surveys for poverty measurement.

[2] In addition, the large and growing gaps between micro and aggregate estimates of household consumption hinder assessments of global progress in poverty reduction and the effect of economic growth on that process (Ravallion, 2003; Deaton, 2005). The extent to which possible measurement flaws in survey design format may contribute to this gap is unknown.

five of the six countries had made a major change in the design of the module collecting consumption data in their surveys.[3] Only South Africa had not changed the consumption survey design across the two years in the database, although they did in 2005. As we will show in this paper, these design changes can result in large changes in both mean consumption and distributional measures.

Our findings come from a study specifically intended to provide more comprehensive evidence on the sensitivity of consumption estimates to varying survey designs. We developed eight alternative consumption questionnaires that were randomly distributed across 4,000 households surveyed in Tanzania. These eight designs all represent common approaches to consumption measurement that have been implemented in numerous settings. We consider the various strengths and weaknesses of each approach to survey-based consumption measurement, including the types of errors that are likely to be more prominent. While in this experiment there are no validation data available for the study households, one resource intensive variant – a personal consumption diary with intensive and frequent supervision – is believed the most accurate and in many ways represents a "gold standard" for field based consumption measurement in survey format. We investigate how alternative survey designs compare with our benchmark, with a focus on the patterns of over or under-reporting of consumption expenditure with respect to household characteristics, and poverty and inequality measures.

The paper is organized as follows. A review of related literature is presented in Section 2. Section 3 describes the experiment, the Survey of Household Welfare and Labour in Tanzania, and construction of the consumption aggregates. Section 4 discusses the results. We compare estimates of total and component consumption for each of the modules, explore how household characteristics affect reporting, and consider the impact of alternative survey designs on inequality and poverty measures. Section 5 discusses the resource requirements for each survey design. The final section reviews the lessons learned herein.

2. Issues in the Measurement of Consumption through Surveys

[3] The variation in the type of consumption survey for these countries included: bounded recall and usual consumption (5 surveys Côte d`Ivoire), bounded recall and usual consumption (3 surveys in Ghana), bounded recall and diary (2 surveys in Malawi), usual consumption and diary (3 surveys in Tanzania), and bounded recall and diary (2 surveys in Zambia). The terms bounded recall, usual, and diary are explained in the next section.

There is a large literature on methods of collecting consumption data through household surveys.[4] Much of the evidence from survey experiments was generated in developed countries, while studies from low-income settings typically draw on surveys not specifically designed to systematically compare contrasting methods. [5] The specific experimental design described in the next section was motivated by the desire to carefully investigate the divergences across the main methods of consumption data collection in common use today in a developing country setting. The primary dimensions in which these methods vary are four: the use of diary vs. recall, the level of aggregation or detail in the commodity list, the reference period and the level of respondent. Some existing evidence of the measurement consequences of the four aspects is reviewed in turn; lengthier discussions of this literature can be found in Gibson (2006), Deaton and Grosh (2000), and Scott and Amenuvegbe (1991).

There are many potential sources of reporting error that lead survey estimates of consumption to deviate from actual consumption. Perhaps the most commonly discussed in the literature is recall error, where a household under-reports true consumption over the period of recall due to faulty memory. Presumably the longer the period of recall, the greater the cognitive demand on the respondent and the greater the divergence between reported and actual consumption. Several studies have documented that, all else equal, the longer the period of recall, the lower the reported consumption per standardized unit of time. Closely related to recall error is telescoping, where a household compresses consumption that occurred over a longer period of time into the reference period asked and thus reports consumption greater than the actual value.

A third important source of error is the inability to accurately capture individual consumption by household members that occurs outside the purview of the survey respondent. Clearly this inability may be more significant for certain types of consumption goods such as transport, telecommunication, or meals outside the home. The degree of inaccuracy likely also increases with the number of adult household members and the diversity of their activities outside the home.

Other sources of error with no obvious direction of bias include rounding error and cognitive errors that result from consideration of hypothetical consumption constructs such as consumption in a "usual" month which may present additional cognitive demands compared to a definitive recall period in the immediate past. All of these sources of error can be considered non-deliberative or unintentional. On the other hand, intentional misreporting can arise if there is perceived social pressure to appear either "wealthier" or "poorer" to the interviewer and thus the social context of the interview may also be

[4] This section does not review the literature from cognitive psychology on the effects of question framing and other issues related to survey design generally; we focus instead on issues and evidence specific to household consumption expenditure.

[5] The developed country studies often compare recall to diary methods (see, for example, Neter, 1970; Neter and Waksburg, 1964; McWhinney and Champion, 1974; Kemsley and Nicholson, 1960; Gieseman, 1987). More recent work examines other dimensions such as bracketing and question wording/prompting, as in Comerford et al. (2009).

relevant. Finally the interviewer and respondent could suffer from fatigue in the latter part of the interview when surveys are long, want to rush to complete the questionnaire, or lack integrity when supervision is limited.

The evidence on the implications of consumption survey design which we review is fragmentary and typically sheds light on only one aspect of survey design while ignoring resource implications of the variants considered. Thus previous work at best either explicitly or implicitly investigates how one source of reporting error varies with design. For example, systematic variation in the length of recall can explore the net relative effect of recall and telescoping errors at different reference periods. A challenge to this previous research is the lack of a consumption benchmark, making it difficult to conclude one design is more accurate than another since there are no data on actual consumption to validate the survey estimates. Scanner data may allow validation in certain contexts in developed countries but will be unavailable for many years in developing countries where many goods are either home produced or bought in informal markets. Prior attempts to form a benchmark for consumption surveys in developing countries have not been fully successful. For example, in India's NSS experiments enumerators visited households every day and gave volumetric containers for measuring food consumption but for some foods less than two-thirds of respondents used the containers and many respondents did not use the daily diary given to them (NSSO 2003).[6]

Diary or Recall

In a non-negligible number of developing countries, including Brazil, China, and many countries of Central Europe and Central Asia (where literacy rates are high), there exists a tradition of diary-based collection of consumption data – at least for certain consumption items such as food. This contrasts with the more common practice in Living Standards Measurement Study (LSMS) surveys and other multi-topic survey instruments to base data collection on recall. It is far from clear to what extent diary and recall surveys are substitutes, or if there are biases introduced by switching from one to the other. In theory, entries are recorded in the diary by the respondent household at the moment of either acquisition or consumption, or soon thereafter. However, in practice it can often be the case that interviewers assist in completing the diary, effectively blurring the line between consumption data collected by diary and by recall interview. This mediation by interviewers is especially likely since most diary surveys instruct

[6] National Accounts (NA) estimates of household food consumption are also not a plausible source of validation data, at least in developing countries where NA data may also suffer from various sources of inaccuracy. For example, in comparisons between survey and NA estimates of food consumption in India, both survey and national account statisticians have concluded that discrepancies more likely reflect errors in the national accounts (Minhas 1988; Kulshreshtha and Kar 2005).

interviewers to return every few days to update the diary in cases where it may not be completed by household members, such as in households with no literate adults.

While well-implemented diary surveys might be expected to yield higher (and presumably closer to actual) levels of consumption, experimental evidence for this from developing countries is fragmentary. In urban Papua New Guinea, Gibson (1999) found that mean total expenditure per capita was 14 percent higher (and food consumption 26 percent higher) using a diary rather than a recall survey. However, preliminary comparisons of consumption from the Bosnia and Herzegovina LSMS (recall) survey with the household budget survey (diary) in 2004 show levels to be similar (World Bank, 2006).

These studies pertain to household-level diaries and thus do not address an important source of mis-measurement – the difficulty for a sole respondent to perfectly capture total household consumption. Personal diaries are generally considered better for obtaining complete household expenditure or consumption data because it is unusual, in most societies, for any one household member to know the expenditure/consumption of every other member, especially on items such as alcohol, tobacco, daily travel, personal toiletries, daily newspapers/magazines, and meals (especially snacks and lunches) eaten outside the home. In Russia, as part of an effort to study the reliability of the Household Budget Survey, a random sample of households in the 3rd quarter of 2003 were assigned personal diaries rather than the household diary. The personal diary yielded expenditure levels which were 6-11 percent higher than a household diary (World Bank, 2005). However, the personal diary was plagued with non-respondent problems in this experiment; about 54 percent of households assigned the personal diary did not complete it. Personal diaries also create the possibility of inadvertent double-counting of items consumed in common but reported separately by two or more individuals, and so care must be taken with this method. We discuss this further in the next section.

While the perfect diary experiment is, essentially, the benchmark for collecting consumption data through survey, it is difficult to extrapolate from experiments between diary and recall consumption modules if the implementation of the diary in the experiment does not reflect the reality of fielding a diary in developing or transition countries. It may be that in the case of illiterate (or simply unmotivated) respondents, a 7-day diary becomes a 7-day recall survey if the information has to be obtained orally by the interviewer at the end of the period. The implications of variation in literacy, motivation, and other factors, although not well-documented, suggest that it can be quite difficult to conduct a high-quality diary survey, regardless of issues related to respondent recall bias.

In their study of measurement errors in food consumption, Ahmed et al. (2006) observe problems in the Canadian Food Expenditure Survey diary related to infrequency and diary exhaustion and conclude that the "superiority of the diary data may not be as obvious as the literature suggests." In Tanzania, the mean number of transactions in the official Household Budget Survey diary declined by 10 percent from the 1st to 3rd survey month, and by 27 percent from the 1st to 13th (last) survey month (NBS, 2002), suggestive of inconsistent supervision of interviewers and/or diary fatigue on the part of respondents. In Malawi, nearly 40 percent of 10,698 households in the 1997/98 national household survey were judged to have incomplete or unreliable expenditure data due to poor maintenance of the diary (National Economic Council, 2000); this drastic waste of survey resources prompted the switch to recall consumption methods in the subsequent national household surveys in 2004 and 2010. In the case of the Russian Household Budget Survey, the pattern of non-response to the household diary, especially among wealthier households, generates extreme sample weights which Gibson and Poduzov (2003) conclude are inefficient from both a statistical and a budget perspective.

Level of Detail in Consumption Questionnaires

The number of items about which data are collected is one of the central issues in designing a consumption survey module. National consumer price index baskets may have in excess of 300 items. To reduce the burden on respondents and survey activities in general, shorter lists of items are created by focusing on those items that represent the greatest share of consumption and aggregating other, less important, items into categories. At present, the number of consumption items (or categories) for which data are collected from households in LSMS surveys ranges from 37 to 305, with the mean being 137 and the median 130: the mean number of food items alone is 75.[7] This level of detail on consumption is thought to be important as it prompts respondents to remember more completely and accurately their consumption. But, the costs may be high: longer interview time, greater respondent fatigue and higher non-response.

While there are surprisingly few studies from developing countries evaluating the accuracy of shorter versus longer lists — none were identified from any African countries — it appears that more disaggregation and longer lists of items result in higher levels of recorded consumption. In El Salvador, data from 1994 showed that the longer, more detailed questions on consumption resulted in an estimate of mean household consumption that was 31 percent higher than that from the condensed version of the questionnaire (Jolliffe, 2001). This resulted in much higher poverty estimates from the short questionnaire, and the geographic distribution of relatively poor persons was significantly different, suggesting some re-ranking of households based on survey design. Pradhan (2001) found that the shorter

[7] The figures come from a LSMS inventory of survey designs. Note that although 100+ consumption items seems a large number, many Household Budget Surveys (HBS) will cover even more food items in the diary. HBSs usually do not collect much additional information on household socioeconomic conditions, whereas the LSMS surveys do, thereby necessitating both recall and food lists that are not exhaustive of all foods. The objective of an LSMS survey is consistent measurement of total consumption for each household in the sample, rather than accurate measures of commodity-specific consumption for the population.

consumption module in the *Susenas* questionnaire in Indonesia results in average consumption that is 12-20 percent below the levels from the longer consumption module, although the ranking of households based on observable characteristics was not different. Earlier experimental work, which led to the introduction of the short-form consumption module in the *Susenas*, also found that a reduction in the number of items from more than 100 to 15 decreased consumption but did so proportionally, thus not re-ranking households (World Bank, 1993). In Jamaica, the shortened consumption modules produced mean per capita consumption results that were about 20 percent lower than the standard modules (Statistical Institute and Planning Institute of Jamaica, 1994). In Ecuador, where two versions of the food module were piloted, the ratio of food expenditures in the long module to the short module was 1.67 (Steele, 1998). Again there appear to be plenty of examples where aggregated commodity lists result in a lower mean and presumably less accurate consumption measure. What is not clear is the relative savings in survey resources, and the effects on distributionally sensitive measures, such as inequality, from such a design.

Reference Period

In recall modules there are also important issues associated with the choice of reference period to be used. The direction of bias introduced by lengthening or shortening the reference period is ambiguous *a priori*. Respondents may have difficulty recalling consumption expenditure with longer reference periods due to diminished capacity to remember (memory lapse). On the other hand, short recall periods may produce over-estimates (telescoping errors) if respondents include consumption/expenditures just outside the reference period. The extent to which expenditures are misreported in these ways is presumed to depend on the item in question and the characteristics of the household.

In India, there has been debate in recent years as to the impact on measured poverty and inequality of having altered the reference period over which (recall-based) consumption data has been collected in two key rounds of the National Sample Survey (see Deaton and Kozel, 2005). Lanjouw (2005) argues that, in Brazil, the application of an inappropriately short reference period over which purchases of foods are recorded in the 2002 *Pesquisa de Orçamientos Familiares* consumption survey diary might account for the surprisingly high frequency of households who report zero consumption of certain key food items.

In Ghana, Scott and Amenuvegbe (1991) ran a well-documented experiment for 144 households on the impact of recall duration on reported expenditure (not consumption) on the 13 most frequently purchased items. They find that each additional day of recall results in about 3 percentage points decline in reported daily expenditure, which plateaued at about 20-25 percent after a recall periods of more than

7-10 days.[8] They attributed this to respondents adopting 'normative' reporting when faced with a recall period that was sufficiently long that the item would have been purchased several times in the period. With a longer period there are too many episodes to count so respondents are presumed to use a rule-of-thumb estimation strategy, the error in which may not be much different between a two week period and a four week period. Other experiments on recall period that were identified are all based on surveys in which the same household reports expenditures across two different recall periods *within* the same survey. This is potentially problematic if there is conditioning bias due to households or interviewers imposing consistency between their responses. [9] Grosh et al. (1995) compare consumption estimates from the Ghana LSMS reported by short (2-week) recall periods and longer recall. The latter were "explicitly normative" questions in which households report the number of months purchases were made in the last year, how often they were made, and how much was usually spent each time (referred to here as "usual-month" questions). While food expenditures were not different, non-food expenditures were 72 percent higher based on short recall, contrary to expectations. Similar analysis is presented for Jamaica (Grosh et al., 1995), Côte d'Ivoire, Vietnam, and Pakistan (Deaton and Grosh, 2000). Deaton and Grosh (2000) conclude that there is only a slight sensitivity to the choice of these recall periods. In much earlier work from Malawi in 1970, Plewis (1972) presents admittedly weak evidence that 7-day food recall is consistent with 24-hours recall (in some sense, this is like a diary). However, it is not clear to what extent conditioning bias is driving these results. The Indian National Sample Survey (NSS) experimented with using a "last week" versus a "last month" recall and found that for the all-food aggregate the estimates based on weekly recall were 21 percent higher (NSSO 2003).

In short there are numerous features of survey design that can result in different reported consumption levels for the same household over the same period of time. Usually based on ad hoc or opportunistic studies, clear divergences arise in measured consumption when recall periods are varied, commodity lists are shortened or lengthened, and consumption is recorded through diary or recall. Underlying these divergences are various reporting errors of likely differing magnitudes, thus making it unclear (i) which variant of consumption measurement more closely hews to the truth; (ii) the implications of survey design for subsequent analysis, especially when consumption measures derived from different variants are combined in the analysis; and (iii) generalizable lessons for future

[8] Dutta Roy and Mabey (1968) present results of studying alternative recall periods in Ghana for experiments circa 1966; they found significantly larger declines in reported expenditure than Scott and Amenuvegbe (1991). They attribute this to a combination of memory lapse for longer recall periods and over stated expenditures for short recall periods, although it is not clear how this conclusion is drawn.

[9] A related research design would be to collect both diary and recall consumption data from the same household. Ahmed et al. (2006) study food expenditure among Canadian households reported by recall (past one month) and then collected for the following 2 weeks by diary. However, this design also risks cross-module spillovers within the household.

consumption measurement in developing country surveys. The current study was designed to shed light on all three questions.

3. Tanzanian Consumption Survey Experiment and Consumption Aggregate

Survey Experiment

Our survey experiment entailed fielding eight alternative consumption questionnaires randomly assigned to 4,000 households in Tanzania. The eight designs vary by method of data capture, level of respondent, length of reference period, number of items in the recall list, and nature of the cognitive task required of the respondent. They are summarized in Table 1. Modules 1-5 are recall designs and modules 6-8 are diaries. These eight designs were strategically selected to reflect the most common methods utilized. The alternative designs focus on variation in the measurement of food consumption and expenditure on select frequently purchased non-food items, and not on general non-food expenditure where, due to the infrequency of purchase, there is a much greater degree of design harmonization in practice. Food consumption includes the quantity consumed from three sources (purchases, home production, and gifts/payments) and, for purchases only, the corresponding value of the quantity (in Tanzania shillings). Modules 1 and 2 seek the consumption values for a long list of 58 commodities. In module 3, the subset list consists of the 17 most important food items that constitute, on average, 77 percent of food consumption expenditure in Tanzania based on the previous Household Budget Survey. When comparing consumption expenditure in module 3 with other modules, we scale up food expenditures for that module (by 1/0.77), as is commonly done in practice. Module 4 is a collapsed list where the 58 food items are aggregated into 11 comprehensive categories.[10]

Among the recall modules, module 5 deviates from a reporting of actual expenditure over a specified time period. Instead it asks for "usual" consumption, following a recommendation in Deaton and Grosh (2000), where households report the number of months in which the food item is usually consumed by the household, the quantity usually consumed, and usual expenditure in those months. These questions aim to measure permanent rather than transitory living standards, without interviewing the same households repeatedly throughout the year. Hence, module 5 introduces two key differences from the other recall modules: a longer time frame and a different cognitive task required of respondents. The longer time frame should allow averaging over short-term fluctuations in living standards, which are caused not just by seasonality but by any short-term idiosyncratic shocks occurring in the survey

[10] Since respondents often use local units for reporting (Capéau and Dercon, 2006), the survey teams also established conversion factors for these local units and surveyed local prices so that quantity and value information in diaries could be checked for outliers.

reference period. Therefore, in comparison with a series of short-period surveys, staggered evenly over the year for an entire sample, there may be less inequality in the "usual" month measure than in formats with shorter recall periods.[11]

Second, the "usual" month format forces respondents to undertake a more difficult cognitive task of *estimation*, rather than just *memory recall and counting* required by other modules. There are many estimation strategies that respondents might try when estimating "usual" month consumption, each with their errors. For example, they could estimate monthly quantity from a rule-of-thumb daily consumption rate ("one mango a day during mango season") and then apply either a lagged price or the current price to convert into consumption values. Or they might estimate monthly spending from mental accounts for shorter periods ("my weekly beer budget is 10,000 shillings, so 43,000 shillings per 'usual' month"). What is most unlikely, however, is that they try to remember all occurrences, count them, and then average them.[12]

The three diary modules are of the "acquisition type." Specifically, they add everything that came into the household through harvests, purchases, gifts, and stock reductions and subtract everything that went out of the household through sales, gifts, and stock increases. For example, items that are purchased to be resold, given away, or kept in stock are not counted as consumption. Two modules are household diaries in which a single diary is used to record all household consumption activities. For the third diary module, each adult member keeps their own diary while children were placed on the diaries of the adults who knew most about their daily activities. [13] Following the literature, we label this as a "personal diary." The personal diary was carefully designed to avoid double counting. Diary entries are specific to an individual and should leave no scope for double-counting purchases or self-produced goods. It is possible that a "gift" could be given to the household and accidentally recorded by two individuals. However, interviewers were trained to cross-check individual diaries for similar items purchased, produced, or

[11] Gibson, Rozelle, and Huang (2003) find the Gini coefficient in urban China is 64 percent higher if respondents kept a diary for one month, with the sample spread evenly across the year, compared with the Gini that results from all respondents keeping the diary for one year. In contrast, the annual mean from extrapolating staggered monthly expenditures almost exactly equals the mean from the annual diaries. Since this experiment was carried out in an urban area, the difference is not primarily due to seasonality but instead to short-term fluctuations within the survey period that are subsequently reversed over the year.

[12] While the literature has findings about errors in memory and counting (more error the longer the period and the more transactions to count), almost nothing is known about errors in estimation strategies. For example, if respondents apply current market prices to usual month quantities, it may reintroduce seasonality into the "usual" month estimate and so will tend to raise inequality toward the level found with short reference periods. In the Vietnam Living Standards Survey, the unit value for rice, formed from the ratio of "usual" month expenditure to "usual" month quantity, has considerable seasonality, with the same pattern across months that is observed in community market price surveys (Gibson, 2007). The community market price surveys reflect current seasonal conditions because they are carried out only once in each cluster but the unit values should not because they are meant to refer to purchases made in a "usual" month (that is, an a-seasonal construct).

[13] Just over 52 percent of individuals in the respondent households maintained a personal diary, with the remaining members (typically children) allocated to a specific adult personal diary.

gifted that occur on the same day and to query these during the checks. In many cases, one person will acquire food for the household (such as buying 5 kilograms of rice), which is entered in the diary of the person acquiring the food. So the personal diary is a not an individual's record of food consumption. Rather, it records the food brought into the household by each member even if for several members to consume (as well as food consumed outside the household). This intensive supervision of the personal diary sample would be impractical for most surveys but these investments were made in order to establish a benchmark for analytic comparisons.

Each of the eight designs varies how food expenditures (including value of home production and consumption) are collected along the lines specified in Table 1. Non-food items are divided into two groups based on frequency of purchase. Frequently purchased items (charcoal, firewood, kerosene/paraffin, matches, candles, lighters, laundry soap, toilet soap, cigarettes, tobacco, cell phone and internet, and transport) were collected by 14-day recall for modules 1-5 and in the 14-day diary for modules 6-8. Non-frequent non-food items (utilities, durables, clothing, health, education, contributions, and other; housing is excluded) are collected by recall identically across all modules at the end of the interview (and at the end of the two-week period for the diaries) and over the identical one or 12-month reference period, depending on the item in question. Any cross-module differences in measured non-frequent non-food consumption we take as due to spillovers from the different amount of memory training or conditioning of respondents that may be induced by the different food consumption modules.

The survey experiment conducted was termed the Survey of Household Welfare and Labour in Tanzania (SHWALITA), and was implemented by a well-established data collection enterprise, Economic Development Initiatives (EDI). This survey was designed and fielded to study the implications of the alternative survey designs for consumption expenditure measures and labor market indicators (see Bardasi et al., 2010 for a description of the labor survey experiment). Here we focus on the component that applies to consumption expenditure measures. There were four alternate labor modules which were each assigned randomly to two-thirds of the households given modules 1-4. We do not find any impact of the inclusion of a labor module on subsequent consumption outcomes for any of the modules with the labor module.

In preparation for the experiment, interviewers were trained extensively and subsequently were under continuous quality control. Regular supervisor re-interviews of households as well as supervisory direct observations were made to prevent any interviewer idiosyncrasies from developing. Each interviewer implemented all eight modules in equal proportion in order not to confound module effects with interviewer effects. There is the possibility of spillover effects in that an interviewer may influence how one module is completed based on experience with another module. Such learning and spillovers may also exist in other surveys, where experienced interviewers have exposure to different consumption modules. To the extent that such an effect may exist, it would be expected to result is an underestimation of any real gap between consumption collected by alternatively designed modules. However, this design was selected by investigators to minimize the presumed more serious risk of interviewer effects confounding the estimated module effects that can occur if individual interviewers were exclusively assigned to one module type.

The experiment used double blind data entry and high levels of quality control to reduce any effect of data entry on estimated cross-module differences. The data entry protocol was the same for all versions of the questionnaire and hence should not be a source of systematic error biasing the comparison of module performance.

The field work was conducted from September 2007 to August 2008 in villages and urban areas from seven districts across Tanzania: one district in the regions of Dodoma, Pwani, Dar es Salaam, Manyara, and Shinyanga and two districts in the Kagera Region. The districts were purposively selected to capture variations between urban and rural areas as well as across other socio-economic dimensions to inform survey design related to labor statistics and consumption expenditure for low-income settings. The sample was constructed to be representative at the district level, but not at the national level. Data from the 2002 Census were used to enumerate all villages in the district. In the first stage of the sampling process, a probability-proportional-to-size (PPS) sample of 24 villages was selected per district. In a second stage, a random sub-village (or enumeration area, EA) was chosen within the village through simple random sampling.[14] In the selected EA, all households were listed. From these lists three households were randomly assigned to each of the eight modules. This was done through simple random sampling starting with module 1 and moving to module 8. The five alternative recall questionnaires were conducted in the span of the 14 days the survey team was in the EA to conduct the household diaries. Fortunately refusal and attrition after starting the survey are not an issue. Among the original households selected for the survey and assigned to a module, there were 13 replacements due to refusals. Three households that started a diary were dropped because they did not complete their final interview. This yields a final sample size of 4,029 households.[15]

[14] The equivalent of a village in urban areas is the *mtaa* (Swahili for "street"). As there is no official subdivision of an *mtaa*, our listing teams carefully defined and delineated Enumeration Areas within the *mtaa* in cooperation with local informants.

[15] We have almost no item non-response (in that the respondent does not report that the household consumed the specific item in the specified period) for food in the recall modules or for non-food in any of the modules. We do not observe any distinct patterns in non-response across our survey designs, or within a recall design by the location of the item on the list.

The basic characteristics of the sampled households generally match the nationally representative estimates from the 2006-07 Household Budget Survey (results not presented here but available from the authors upon request). Household interviews were conducted over a 12-month period, but because of relatively small samples within the period, we do not explore the survey assignment effects across seasons. Instead, the analysis reports the mean effect of questionnaire design across all seasons in Tanzania since each season is equally weighted in the data.

The randomized assignment of households to different questionnaire variants appears to have been successful in terms of balance across characteristics relevant for consumption and consumption measurement when examining a set of core household characteristics, presented in Appendix Table 1. The table presents these mean characteristics by each module. At any point where there is a significantly different pairwise comparison, those pairs are indicated. Out of a possible 420 pairwise comparisons in Appendix Table 1, only 13 pairs are significantly different at the 5 percent level.

Consumption Aggregate

We construct annual consumption aggregates (total household consumption expenditure) consistent with standard practices (see Deaton and Zaidi, 2002) but adapted to module specifications. For modules with the quantity of own-produced consumption but not shilling values (modules 1-3), monetary values were computed from unit values constructed from household information on purchases.[16] In each of the three modules, about 50 percent of quantities were able to be transformed by household-specific unit values (i.e., households also purchased the item), while in 25 percent of the cases the unit in which the household reported own-produced consumption was Tanzanian shillings. In the diaries, for each item we take the quantities reported minus what was recorded as sold or given away (deducted from the value as a percentage). Non-item specific (lump sum) corrections were made for changes in stocks (added consumption from stocks and adjusted for recorded acquisitions that were stocked instead of consumed), for food given to animals, and for items that were recorded as "forgotten".[17]

[16] For non-purchased products, we used the first available unit price in the following order: (i) Household-specific unit value, only available if household also purchased item (values more than five standard deviations from the district median price were replaced with district medians); (ii) EA median unit value; (iii) District median unit value; (iv) Overall median unit value; (v) EA median price from price questionnaire; (vi) District median from price questionnaire; (vii) Overall median from price questionnaire; and (viii) A price assigned by EDI staff, using best guesses based on the prices of similar items or units in the EA or district - only a handful of consumed items were valued in this way.

[17] Forgotten items refer to the following. At the end of each of the 2 weeks of the three diary variants, the household is asked about the three main foods eaten daily and then the diary is reviewed to ensure that acquisition of these foods (including from stocks) are recorded. If these foods were omitted in the diary by mistake, the information is then collected by the enumerator (effectively in the form of recall).

Module 3 stands out since it explicitly does not cover the same range of food as the other modules. Rather it is a subset list of the most frequently consumed foods based on the Household Budget Survey data. We scale food totals measured with this module up by 29.87 percent in order to compare with totals from the other modules, which are all covering a broader universe of foods (as stated earlier, we base this scaling up on the Household Budget Survey, where this subset of goods accounts for 77 percent [=1/(1+0.2987)] of total food consumption). For every module, we exclude expenditures on taxes and contributions to ceremonies. These consumption categories are not typically included in consumption aggregates because they do not add to utility and are idiosyncratic and infrequent by nature.

The direct comparison of results from the different survey modules sheds light on the relative influences of the various reporting errors discussed in Section 2. Since the frequently supervised personal diary is believed the most accurate in that it minimizes the magnitude of recall lapse, telescoping, and missed individual consumption, comparisons with this benchmark will be of particular use in understanding the net impact of the various reporting errors in each of the different modules. However, comparisons across variants other than the benchmark will also be informative for the performance of each module. As discussed, every attempt was made to ensure the personal diary is close to 'true' consumption and as a result there appears to be very little double-counting or consequences of respondent fatigue. We find no differences in the number of diary entries and the total consumption expenditure between the first and second week, either overall or by key household characteristics.

4. Results

Differences in Mean Consumption Estimates

We begin our examination of the alterative consumption modules by investigating differences in reported per capita expenditure across modules. Table 2 presents the means and medians of per capita consumption by module for total, food, frequent non-food (either collected by recall or diary), and non-frequent non-food (all recall) consumption. The other seven modules generally report lower consumption, especially food consumption, than the personal diary (module 8) benchmark. The module that comes closest in both mean and median to the personal diary is module 2 – the 7-day long list module. Module 4 (7-day collapsed list) has the lowest mean and median levels of food consumption.

The simple comparison of means already reveals some patterns that will persist in the analysis to follow. For example, the inability of household-based diaries to fully account for individual consumption, including that outside the household, contributes to 19 percent lower mean consumption and 15 percent lower median per capita consumption in the study sample. This is from a direct comparison of module 8 (personal diary) with module 6 (household diary supervised at the same frequency). In the same comparison, food consumption is 22 percent lower while total non-food is 16 percent lower (at the mean). Perhaps somewhat surprisingly, in the less-frequently supervised household diary the reported food consumption is 8 percent higher than in the more regularly supervised household diary.

When looking at the recall modules, the 7-day long list recall (module 2) has the highest mean food consumption, even slightly higher than the personal diary, followed by the 14-day recall (module 1) and then the 7-day subset list recall (module 3) after it is re-scaled. As expected, a full listing of commodities results in a greater total value of consumption, in fact a 21 percent increase when comparing module 2 with the collapsed 7-day recall (module 4). This substantial difference indicates that the decision to adopt aggregate (shorter) consumption lists should consider not only the extent of savings in survey cost or reduced respondent fatigue, but also the potential downward bias in measured consumption. In terms of recall period, the 7 percent decline in reported (annualized) food expenditures as we move from a 7 to a 14-day recall period is consistent with other studies, but we should not take this to mean that 7-day recall is necessarily the more accurate method since the higher values may simply reflect differing magnitudes of recall and telescoping error.

The recall module that asks about the "usual" month consumption for the last 12 months tends to report food consumption that is lower than the 14-day recall.[18] This is not surprising since with the lengthened recall period a greater degree of recall error is expected. However, we cannot separately identify this cause of divergence from the other deviation in design, which is to ask about the "usual" month rather than a specific recall period. One indication that the "usual" period approach does result in response differences is found when looking at non-food expenditures for which there is no variation in questions across recall modules. Frequent non-food expenditures are much higher for the "usual month" respondents, on average 25 percent higher than with the full list 7-day recall (module 2) and 32 percent higher than the aggregated 7-day recall (module 4). Given the much longer time to complete module 5 for food (discussed below) which would suggest respondent fatigue, the higher non-food amounts reported in module 5 is surprising. It suggests that it is not fatigue but, rather, the cognitive demands of food recall for a "usual month" in module 5 which affects subsequent responses.

For a formal test of difference across module designs we regress the natural logarithm of food, non-food, and total consumption on dummies indicating the module assignment (with the personal diary as the left-out category unless otherwise mentioned). The log-specification allows us to interpret the

[18] As stated earlier, the survey was conducted over 12 months in a balanced fashion, so any differences between Module 5 and other recall formats are not due to seasonal effects.

coefficients as approximate percent deviations in mean value from the excluded category. Because the survey experiment was randomized, we do not include any controls.[19]

Formally, the framework is as follows:

(1) $$C_{ik} = \beta_k M_k + e_{ik}$$

where C_{ik} is (log) consumption of household i assessed with questionnaire k (k = 1… 7). M_k is a vector of dummy variables for module type. Randomization of modules assures that the residual error term, e_{ik}, is orthogonal to M_k. Consumption refers to total consumption or its three subcomponents (food, frequent non-food, non-frequent non-food). The expenditure level from questionnaire k in relation to the excluded category, module 8, will be given by β_k, which can be compared with the relative cost of administering that questionnaire type. Additional analysis that explores how consumption reporting varies by particular household characteristics such as household size or education of household head adopts the related specification:

(2) $$C_{ik} = \beta_k M_k + \beta_x X_{ik} + \gamma_k M_k X_{ik} + e_{ik}$$

where X_{ik} is a single household characteristic and γ_k reports the coefficient on the interaction term. We estimate Equation (2) separately for each of the selected household characteristics.

The regressions in Table 3 compare each of the seven modules against the personal diary and confirm the results from the first descriptive table. [20] The results in column (1) show that, with the exception of 7-day recall with the long list, other modules record between 7 and 28 percent less consumption compared with the personal diary. The impact on food consumption is of a similar magnitude (column 2). Having just one respondent complete the diary for an entire household is associated with significantly lower consumption of 14-17 percent, potentially because unobservable personal consumption is not explicitly captured in the design. Differences in frequent non-food expenditures are also observed, especially in the diaries. Because the questionnaire wording and structure for the non-frequent non-food consumption section was identical across all 8 modules, it is perhaps surprising to see significantly negative coefficients for the module 1, 4, and 7 dummies. Such differences can result from three sources: respondent fatigue, since these recalled items come after lengthy food recall sections in modules 1-5 or after a two-week diary; cognitive framing; and differing ability to capture personal non-frequent non-food consumption outside the purview of the main respondent. Contrary to concerns of respondent fatigue, module 4 with the collapsed food categories and shorter interview time yielded significantly less (14 percent less) non-frequent non-food consumption. Possibly the lack of

[19] The addition of controls may introduce bias into the experimental estimator (Freedman, 2008). Nevertheless, if we also control for the demographic composition (detailed age-sex categories), EA fixed effects, interviewer fixed-effects, and other household characteristics, the results are almost identical to those reported.

[20] If, instead of OLS, we estimate quantile regressions at the 35th percentile, a common cut-off for poverty analysis, the results are largely similar to Table 3 (available upon request).

follow-up during the diary period made the module 7 respondents less diligent in the non-frequent non-food section of their final interview.

To focus more sharply on the consequences of design differences among recall modules, in Table 4 we compare modules with different recall periods but the same detail of the food list (modules 1, 2, and 5 in panel A) and modules that have varying detail of food lists but the same recall period (modules 2, 3, and 4 in panel B). For all five of these modules, the non-food expenditure questions are identical (in detail and recall period). In panel A, columns 1-4, we see that lower total consumption associated with the longer recall periods is mainly driven by lower food expenditures. In the case of the "usual" food module, frequent non-food is statistically significantly higher (16 percent) than the 7-day module, and the non-frequent non-food is also higher (8 percent) but this difference is not precisely estimated. Given that the non-food questions are identical, we attribute this difference to the adaptation of respondents to the cognitive demands of the "usual" food questions, which immediately preceded the non-food questions.

Comparing the differing length of food lists (panel B columns 1-4), the collapsed list results in lower total consumption as a result of lower measured food consumption. Reported food consumption is almost 32 percent lower while the subset list, when scaled, results in only 6 percent lower consumption. There are no statistically significant differences in either of the non-food categories, indicating no observable effect of respondent fatigue from the detailed food list on subsequent non-food categories. Clearly, at least in the context of Tanzania, the 17 most important food items (a subset of the full list of 58) performs only marginally worse than the full detailed listing.

Table 5 compares results among the three diaries. In Panel A, columns 1 and 2, we find that household diaries record significantly less total and food consumption than the personal diary, from 13 to 20 percent lower. Between the two types of household diaries, the frequency of interviewer visits makes little difference in total consumption, but food consumption is 7 percent higher in the infrequent diaries than the frequent household diary (and this difference is marginally significant at the 10 percent level; results available upon request), suggesting that the net impact of telescoping and recall error in the infrequently supervised households is positive but not especially large. Interestingly, the frequent non-food consumption exhibits the largest deviations in the household diaries from the individual diary, suggesting that a relatively large amount of these goods are consumed outside the purview of the household respondent.

In panel B of Table 5, we subdivide the household diaries on the basis of the literacy status of the household. Approximately 16 percent of households had no literate adults to fill in diaries. Households assigned to the diary with infrequent field visits had more visits if the household was deemed illiterate (see Table 1), although not as many as for households in the frequent visits sub-sample (module 6). The presence of an illiterate member does not affect reported consumption when households are regularly supervised (panel B columns 1-4, module 6). However, infrequent supervision results in significantly lower reported consumption among illiterate households, even for non-food non-frequent items, which are asked only in recall format. Clearly, leaving a diary with a household lacking a literate member will yield greater mis-measurement unless this difficulty can be overcome through frequent visits, thus converting the survey format to a de facto high frequency recall.

Consumption Measurement and Household Characteristics

The literacy of household members is one characteristic that may affect the relative performance across modules. Table 6 explores the influence of additional characteristics by estimating equation (2) for the following sequence of characteristics: total household size, the number of adult household members (age 15 and above), urban location, education of the household head, and an asset index as an alternative measure of household resources (asset wealth) derived from housing conditions and household durable goods (Filmer and Pritchett, 2001). All of these characteristics are strongly related to true household consumption, but also can affect the accuracy of consumption reporting. For example, respondents with more years of education may be more able to accurately calculate household consumption over the recent past, or more able to estimate consumption in a "usual" month abstracted from consumption in any actual month. Respondents in urban areas or asset-wealthy households may experience a greater variety of consumption in any given reference period and so accurate recall may present additional challenges. Due to space constraints, the results in Table 6 omit level effects and standard errors (results available upon request). The significance of the interaction terms is indicated by asterisks.

Household size is a significant determinant of overall household consumption (larger households typically have lower per capita consumption), but Table 6 also reveals it to be a determinant of increasing divergence from benchmark consumption in select recall modules. For three out of five recall modules (modules 1, 3, and 5) the discordancy in measured consumption increases with the number of household members. The interaction terms for the other 2 recall modules are also negative but not precisely estimated. The same pattern holds when we restrict to food consumption, and grows in magnitude when looking only at frequent non-food expenditure (and for this category the interaction terms for all recall modules are now significant). This may be due to increased cognitive demands of one respondent asked to recall the consumption of an entire household as the number of members increase. Alternatively the relative importance of out-of-household consumption may increase with household size, particularly for frequent non-food expenditure, which is often privately consumed (e.g., cigarettes, cell-phone top-ups, bus and taxi fares) and this consumption is systematically missed by reliance on a single household respondent. However, if the lone respondent misses some amount of personal consumption, it appears to be a factor only for recall modules. The household size interaction terms for either household diary module are not significantly different from zero.

The second column of each panel in Table 6 refines the above discussion by looking at the influence of the number of adult (age 15 and above) household members. The effect of household size on reporting accuracy is strengthened when looking only at the adults in the house. With regard to overall consumption, the interaction terms for four recall modules are now significantly negative (all but module 4) and are at least twice the magnitude of the coefficients on total household size. The same general results hold for food consumption and, especially, frequent non-food purchases, suggesting perhaps that it is largely adult consumption that is missed by sole respondents, either because this consumption takes place beyond the purview of the respondent or adult consumption is more complex than child consumption (more of which is food from a common pot), and more difficult to recall. The interaction coefficients for the household diaries are also negative but not significant. These diary coefficients are also smaller in magnitude, suggesting again that performance with respect to household size differs dramatically by recall or diary format.

Urban households with recall modules do not perform appreciably worse than their rural counterparts vis-à-vis the benchmark. No interaction term for urban location is significant for any recall module except module 5, the "usual" month recall. For this module, the difference in total and food consumption from the benchmark is attenuated in urban areas, suggesting that the cognitive demands of the usual month construct present particular challenges for rural households. This may be because urban households rely more on the market for their consumption, with potentially smoother prices and availabilities over time so that rule-of-thumb extrapolations from daily consumption to a "usual" month are more accurate. Alternatively, this may reflect the higher education of urban respondents since in the results that control for the education of the household head, the only significant interaction term with years of education is again for module 5.

Education of the household head does not affect the performance of the household diaries relative to the individual diary but urbanity does, with household diaries performing significantly closer to the benchmark in rural areas. In fact the lower mean consumption of the infrequent household diary (module 7) summarized in panel A of Table 5 is entirely due to urban households: the level effect of module 7 is insignificant (result not shown in Table 6). In other words, there is no significant consumption difference among rural households that were administered module 7 or module 8. However, urban households with module 7 report 29 percent lower total consumption, 26 percent lower food consumption, 45 percent lower frequent non-food, and 31 percent lower non-frequent non-food. The frequently supervised household diary (module 6) also reports significantly lower non-food consumption (of either category) when administered to urban households. These results are consistent with the supposition that urban settings provide more diverse personal consumption choices (such as outside meals and public transport) than rural areas, and much of this consumption, out of what Deaton and Grosh (2000) term "walking around money", is missed by the sole diary keepers of modules 6 and 7.[21] The difficulty in capturing this "walking around money" may be somewhat mitigated by the frequency of supervision since the magnitude of the interaction terms are less for the frequently supervised module 6 and indeed the interaction term for food consumption for that module is not significantly different from zero.

The final characteristic in Table 6 is a measure of the asset wealth of the household. We see that consumption reporting with respect to the benchmark diverges dramatically depending on whether we look at the recall or diary format. All recall formats have a significant positive interaction term with the household asset index, implying that asset-poor households significantly under-report recall consumption vis-à-vis asset-poor households with the benchmark module, while asset-rich households report roughly equal or even greater amounts of consumption than asset-rich benchmark households. The asset index is a normalized mean-zero standard deviation-one random variable, which implies that rich households with an index score of 1 to 2 achieve consumption parity with rich benchmark households, depending on the particular module administered and the estimated level effect. This finding suggests that net recall error may be particularly acute among poorer households. If this is the case, then recall modules may actually overstate the degree of consumption inequality in a population relative to actual consumption. This possibility is explored further below.

Household diary performance interacted with the household asset index yields qualitatively different findings from that for recall, although this conclusion is more tentative given the relative lack of estimate precision. For both household diary types, the interaction terms are negative, and statistically significant for the infrequently supervised diary for select consumption sub-

[21] This supposition is supported by the differential number of line item entries in urban and rural diaries. While the mean total number of unique commodities consumed by the household over the 2-week period is roughly equal across diary formats in rural areas (ranging from 108 to 111 items depending on the module), the personal diary in urban areas records significantly more commodity types consumed than either household diary. In urban areas, the infrequent household diarists record consumption of 139 unique commodity types while the frequent household diarists record 140 commodities. In contrast, households with personal diaries report consumption of 158 separate commodities over the same 2-week period.

components. Household diaries diverge from actual consumption as the wealth of the household increases, likely due to the increasing importance of private information about consumption as households grow wealthier. Where recall surveys may overstate the degree of inequality in the population, household diaries may actually understate it.

Distributional Measures

The results so far have focused on mean differences in measured consumption but it is possible that module type also affects distributional measures, as suggested in the sub-section above. Table 7 reports the Gini coefficient and its jackknife standard error for each module over the four consumption categories. When looking at total per capita consumption, inequality for the diary-based modules is lower than for the recall modules, although the differences are not always statistically significant. The same general pattern persists when looking at food consumption and the differences between recall and diaries are especially large for frequent non-food consumption. Inequality differences for non-frequent non-food consumption items are relatively slight as we would expect, although the recall modules still yield somewhat more inequality for this type of consumption. [22]

One surprising result from Table 7 is that the highest measured inequality is for the sample given the "usual" month recall. The aim of the "usual" month format is to measure consumption over a longer time frame, averaging over short-term fluctuations in living standards to yield lower inequality. Hence finding higher inequality with this design suggests that another effect is operating and a likely candidate is the heavy cognitive burden of this design. As noted below, the interviews for the "usual" month module took far longer than for any other interview, and the understatement by this module is most apparent for the least educated households. So it may be that some degree of inequality in education (the Gini for years of schooling of household heads is 0.43) is combined with the actual consumption inequality when a cognitively burdensome module like the "usual" month format is used.

Among the other recall modules, the subset and collapsed list (modules 3 and 4) generally yield a more equal distribution. For both module types, this may not be surprising since the relative lack of prompts among the aggregate categories is expected to yield compressed consumption measures due to omissions. If the diversity of consumption goods is greater among wealthier households, this error type will lead to truncated distributions at the upper end. Looking within the

[22] The pattern of results also holds for Generalized Inequality measures GE(0), GE(1), and GE(2) (not presented but available upon request).

diary formats, the personal diary (module 8) yields slightly higher inequality, suggesting that the extent of omission of individual consumption in household diaries is not constant across wealth levels (as we have seen in Table 6 with respect to urban/rural differences and asset-index values) and this variation results in more compressed distributions.

Should these inter-module differences in Gini coefficients in Table 7 be considered large? The difference between the Gini coefficient calculated with the benchmark diary and with module 1 is 14 percent and with module 5 it is 19 percent. These are equivalent to the average gap between consumption-based and income-based Gini coefficients reported by Deininger and Squire (1996), the basis for the current UNU/WIDER WIID database. Thus, just as those authors were careful to identify which inequality estimates came from consumption surveys and which from income surveys so that investigators might adjust them to a consistent basis, so too would it be worth identifying which Gini coefficients are computed from surveys that use diaries and which recall, and over what reference period and level of respondent. Yet such meta-data are hardly ever reported, forcing investigators who combine inequality measures from various surveys to introduce considerable measurement error into their analyses.

Implications for Poverty Analysis

The fact that both mean consumption and distributional measures vary substantially across the modules implies that there will be complex comparability problems when poverty estimates from different modes of consumption measurement are combined. For example, consider a poverty monitoring survey that switched from 14-day to 7-day recall. This switch would raise reported mean consumption (see Table 3) and lower measured inequality (see Table 7). In a setting where the poverty line is set below median consumption, these effects will work in concert to make poverty look lower than it would if the method of consumption measurement had not changed (as occurred with changes in India's NSS). Conversely, if a switch was made from a disaggregated recall list to a collapsed list (as the Indonesian *Susenas* does every 3rd year), it would lower the mean (implying higher measured poverty) and lower inequality (lower measured poverty) with ambiguous implications for poverty measures. Hence there are unlikely to be simple correction factors that can be used to enforce comparability on different consumption modules when attempting to measure poverty consistently over either time or space. Similarly, the literature that reacts to possible survey understatement by combining survey measures of inequality with national accounts measures of mean consumption (such as Bhalla, 2000, and Sala-i-Martin and Pinkovskiy, 2010) is likely to be misguided since survey design changes will also affect estimated inequality.

We attempt to quantify the implications of survey design for poverty analysis in a setting like Tanzania by presenting aggregate poverty numbers based on an international poverty line, as well as the characteristics of those households deemed poor by each module. Table 8 presents standard poverty measures based on a common international poverty line of $1.25/person/day converted into 2008 Tanzanian shillings on a PPP basis, as well as head count poverty estimates based on an alternative poverty line of $0.78 that fixes a 20% poverty rate for module 8 households.[23] Results with this second poverty line are included for comparative purposes.

With the $1.25/person/day absolute poverty line, we see dramatically different headcount rates depending on how consumption is measured. The benchmark personal diary records the lowest level of poverty at 47.5 percent of the population, followed by the 7-day long list (module 2) at 54.9 percent. Even though module 2 yielded mean consumption levels closest to the benchmark (Table 3 lists consumption at 3.9 percent lower and not significantly different from benchmark consumption), the fact that module 2 inequality was also higher (the estimated Gini is three points higher – Table 7) results in substantially higher poverty numbers. The 7-day subset design (module 3) reports poverty almost identical to module 2 – 55.1 percent – while the other recall modules report poverty headcounts all over 60 percent, with module 4, the collapsed list, yielding a poverty estimate of 66.8 percent. Poverty numbers derived from household diary figures also record substantially higher poverty levels than the benchmark – the infrequent household diary records the poverty rate at 55.6 percent and the frequent diary at 59.5 percent.

The poverty headcount is attractive for its ease of understanding, but is also notorious for its implied welfare discontinuity at the poverty line as well as the inability to measure the extent or depth of poverty (Ravallion, 1996). The second and third columns in Table 8 present two additional poverty measures, the poverty gap and the squared gap measures, that give greater weight to the poorest households among those below the poverty line. The overall message is the same. The lowest poverty figures are for the benchmark module 8 while the highest poverty measures are observed in recall modules, especially modules 1, 4, and 5. The ranking of poverty scores varies somewhat by poverty measure. While the 7-day full recall module 2 had the second lowest headcount estimate (54.9%) followed by module 3 (55.1%) and then module 7 (55.6%), module 7 has the lowest poverty gap measure (18.9%) followed by module 2 (19.1%) and then module 3 (21.4%). The largest proportional differences across module are apparent in the squared gap measure where the highest estimated squared poverty gaps, those for modules 4 and 5, are more than twice as large as the benchmark squared gap measure. The changes in module rankings

[23] Remember our consumption aggregate does not include housing so actual reported poverty for 2008 Tanzania should be lower than the numbers presented in this exercise.

across the different poverty measures, as well as the greater proportional differences in the squared gap measures compared with the poverty headcount, reflect the differing degrees of inequality below the poverty line measured by the various modules.

The same divergent poverty estimates across modules are apparent with the lower poverty line of $0.78/person/day, both for the headcount as well as other measures not reported (results available upon request). The differences across modules are not unique to one particular poverty line. In fact the proportional differences in the poverty headcount are even greater for this lower poverty line because the lower poverty line estimates in turn are more sensitive to inequality. Further, and for both poverty lines, the vast majority of the differences between module poverty estimates and the benchmark are significantly different at standard levels of precision. The two modules that, while still reporting higher poverty estimates than the benchmark, are not significantly different are modules 2 and 7.

Clearly poverty measures are highly sensitive to the mode of consumption measurement and care must be taken when poverty studies combine measures of consumption poverty, either across countries or within a country but over time, derived from differing survey designs. Another important goal of poverty measurement, however, is to identify the characteristics of poor households in order to inform the targeting of social policy or to better understand poverty determinants. Hence an equally important consideration, other than whether survey design influences estimated poverty levels, is whether the mode of survey design influences the identified characteristics of the poor. We approach this question through a simple regression framework similar to equation (2) where, using a linear probability model, we regress the household specific poverty indicator on module type interacted with household characteristics. The results, in Panel A of Table 9, again adopt the $1.25/person/day line used in Table 8. A complementary approach to the same general question of the relative rankings of households is to utilize the full information of the continuous consumption distribution, as opposed to the binary poverty indicator, and regress the percentile of the household in the consumption distribution on module type and select characteristics. These rank regressions are presented in Panel B of Table 9 (here again the standard errors and level effects are suppressed due to space constraints but available from the authors).

The results from the analysis in Panel A show virtually no difference in the characteristics of poor households identified by any module. The number of adults in the household, the rural/urban location, and the education of the household do not predict greater or lesser likelihood of household poverty in any module with respect to the benchmark measure – while all three characteristics are important predictors of poverty in their own right, none of them affect the relative probability of being deemed poor across the modules. Total household size is a significant relative predictor for two modules – larger households are significantly more likely to be deemed poor when administered the long 7-day recall (module 2) while larger households are significantly less likely to be deemed poor when administered the frequent household diary (module 6). These estimated effects are not especially large in magnitude – the addition of a household member affects the poverty indicator probability by 2 percentage points in either direction depending on the module. The only other characteristic that is significant at standard levels is the asset-index interacted with the usual month method (module 5). An increment in the asset-index of one standard deviation reduces the relative likelihood of poverty indication by 6 percentage points for module 5 households. Thus a relatively large change in household assets changes the relative likelihood of poverty inclusion only slightly.[24]

Turning to the rankings of household consumption in Panel B of Table 9, we observe a very similar pattern to Panel A. Few observable characteristics result in different within module rankings. Neither the education of the household head or the household asset index affects the relative rankings of households for any module vis-à-vis the benchmark. Overall household size does have a moderate affect on household percentile score for module 3 households, where an increase in household size of one member reduces the ranking of the household by one percentile point. The number of adult household members affects the ranking of 7-day and 14-day full list recall, where an additional adult in modules 1 and 2 households lowers the relative household rankings by 3 percentiles. Finally urban households are ranked an average of 8 percentiles lower in module 7 households than the benchmark. All other interaction terms are not significantly different from zero. As we have seen in this sub-section, while absolute poverty measures vary dramatically across the different modules, the partial correlations of poverty status with selected characteristics do not differ significantly and the relative rankings of households with those same characteristics are largely stable between modules.

5. Resource Implications

[24] An alternative relative approach to poverty measurement sets identical 35 percent poverty headcounts across modules. Even though the poverty rate for each module is fixed at the same level, the characteristics of households below that cut-off can, in principle, differ. Using the same analytic framework as Panel A, however, reveals virtually no interaction term to be significant when the indicator is relative poverty. Hence whether poverty is defined in absolute or relative terms, poverty assignation is largely consistent across modules with respect to observable household characteristics. The same conclusions hold when either the absolute or relative poverty indicator is regressed on module and characteristic interactions with a binary dependent variable model (probit or logit) rather than a linear probability model.

The willingness to accept some degree of inaccuracy in consumption estimates from using a recall survey or a household diary compared with the personal diary is driven by the practical need to reduce the study costs and the burden to respondents of survey work. Our benchmark of frequently supervised individual diaries is often not a feasible option for field researchers, which leads to the question of recommendations for alternative approaches. Here we review the cost implications of different survey designs, focusing first on the cost in terms of time to administer a recall module and then on dollar expenditures for survey implementation in order to contrast recall with diary approaches.

Among the recall modules, the shorter lists in the subset and collapsed recall modules are assumed to significantly reduce the length of time to complete the module and this consideration in our experiences is often used as a justification to adopt a shorter consumption module. Figure 1 presents the average time to complete the recall modules. Respondents required an average of 50 minutes to complete the recall module that listed 58 foods for a 14-day reference period; for the 7-day full list module, the length of time to complete is just 1 minute shorter. The mean time savings is only 8 minutes (42 minutes total) when using the 11 broad food group headings (module 4) and 9 minutes (41 minutes total) when using the module 3 subset recall list with the 17 most important food items. The lack of time saved despite reducing the length considerably reflects the lack of dietary diversity in this setting; the 17 items in module 3 were selected as to omit the less common foods. While the interview time is shorter on average, the 8 or 9 minute savings in interview time of the shortened modules must be weighed against the overall lower mean consumption (especially for module 4) and truncated distribution relative to the longer list recall that these two survey modules generate. These results, of course, pertain to the study setting and in Tanzania households do not have a great deal of dietary diversity. In other settings with greater diversity, these estimated relative time costs may be less applicable.

The most demanding recall module, in terms of the average time taken by respondents, was by far the "usual" month consumption module. The module requires respondents to think about the number of months they consume particular food items and the typical consumption in those months. This is evidently a time-consuming computational task – taking 76 minutes on average. The previous sections have demonstrated that this method results in consumption measures quite divergent from the "benchmark" and may also significantly over-estimate the degree of inequality in the distribution. Given both its poor relative performance and onerous time imposition, this method is not a recommended choice at least for this study setting.

Turning to dollar costs as a way to assess the resource utilization of the different diary modules with respect to a recall survey, the exact costs involved to administer each of the questionnaires and diaries depends on a number of context-specific factors, such as the price of labor, population literacy levels, costs of travel between the sample EAs, printing costs, overheads, and so on. The main drivers of cost differences, however, are relatively easy to pinpoint. They are (i) differential interviewer-days needed to complete a household, and (ii) differences in time needed to enter the data into electronic format. There are no notable differences in fixed set-up costs, so we only discuss the variable (per questionnaire) costs below. However, ignoring the fixed costs of survey implementation means we are over-stating the cost differences between approaches.

Table 10 summarizes these differences and their cost implications under a set of assumptions about time to walk between households, review diaries, and implement recall modules. We are not considering a large multi-topic survey, but a short questionnaire with a complete consumption module (either by diary or recall). In the most labor-intensive set-up, for an interviewer doing personal diaries with daily visits, six households can be interviewed in 17 days. The alternatives to the personal diary are household diaries with varying visits in combination with a second enumeration area (to avoid days with no work in the enumeration area, given that the duration of the diary is for 14 days). For the analysis in Table 10, no distinction is made between the different types of recall questionnaires as they all yielded around four interviews per interviewer per day, which is three times less intensive than in the 'light' fieldwork set-up of household diaries with infrequent visits.

Once the diaries and questionnaires come back from the field, they go through several steps before being available in electronic format: administration and filing of forms, coding of items, double blind entry, reconciliation of the two entries, and so forth. We use information from the survey experiment to compute the time for these activities. On a given day, a data entry operator can process 2.86 personal diary households, 4.31 household diary households, or 8.51 recall questionnaire households.

Based on the person days for field work and data processing and with data entry staff in the office at half the cost of an interviewer in the field, we then compute the total costs of field work (Table 10, column 6). With the average variable cost to administer a recall questionnaire set at US$100 per household (the numeraire to scale the other field costs), the variable cost of the personal diary will fall between US$597 and US$974, depending on whether the household is visited every day or every two days. For a frequently visited household diary, the estimated variable cost lies between US$442 and US$726; for the infrequently visited household diary, it will lie between US$280 and US$334. Clearly the diary approach, especially the benchmark of a personal diary, is a much more resource intensive form of consumption measurement, roughly six to ten times the cost of administering a recall survey to the same household. The cost-savings from a frequently supervised household diary is a relatively modest 25 percent of total personal diary costs. Given the performance of the household diary, especially in urban areas, these resource savings may not be worth the trade-off in terms of downward biases in consumption measurement.

6. Conclusions

In most developing countries, consumption expenditures measured by household surveys are the empirical basis for studying poverty and inequality. Yet there is substantial variation both within and across countries in the design of consumption modules. These differences result in challenges to the cross-population comparisons of poverty and living standards and to efforts to study changes in poverty over time. This study explores the implications of alternative consumption modules design in Tanzania. We designed and fielded eight alternative surveys, selected to capture the most common designs in practice. We benchmark our comparisons to frequently-supervised personal diaries, which we assume come closest to measuring actual consumption by accounting for personal outside-the-house consumption as well as minimizing recall and telescoping error due to the frequent supervision. This method is, however, impractical for large-scale survey work due to high variable costs – we estimate these to be roughly 6 to 10 times as much as for a recall format, and roughly twice as much as for infrequently supervised household diaries.

With regard to diaries, the quality of reporting in household diaries did not vary a great deal with the frequency of field staff visits designed to minimize recall and telescoping error. This is true for all consumption categories (although food consumption was marginally higher (6 percent higher) in the infrequent diary). One notable exception is the literacy of the household. The diary method will dramatically underestimate consumption if the household is illiterate and receives infrequent supervision. For other households, the relative similarities in measured consumption between the frequently and infrequently supervised diary modules suggests that resources can be saved through the infrequent supervision schedule. The gap between both household diary types and the personal diary likely reflects the omission of personal and out-of-household consumption in the household diary, and this gap, while not noticeable in rural areas, is substantial among urban households. The design decision is then whether to invest in the full personal diary approach, which costs roughly three times as much as the infrequent household diary, or cover many more households with the infrequent household diary. Clearly the characteristics of the study population will be a critical factor in this decision – if the population contains a large proportion of urban households, the need for extra resources to implement a personal diary may well be justified.

From among the recall surveys, all far less expensive than diaries, certain recommendations are clear. For one, the savings in survey time from a reduced number of consumption categories in the recall list through aggregation (the collapsed list module 4) is minimal compared with a substantial cost in terms of loss in accuracy. We do not recommend this module type. On the other hand, the other short commodity list design (the subset list module 3), when scaled up based on reference data, performs very close to the longer list form and may be a suitable substitute to longer list recall modules. The gain from such a reduction in list length, however, is slight – less than 10 minutes of interview time – and researchers would have to decide whether the loss of additional detail in consumption information is worth the moderate reductions in interview time. The hypothetical "usual" month recall almost doubles the interview time while reducing the accuracy of measured consumption leaving no practical basis to recommend this approach.

These considerations would lead to a recommendation of a long-list recall module with a reference period of 1 or 2 weeks. Both modules take the same amount of time to implement. The 7-day recall results in mean consumption closest to the benchmark as well as summary inequality measures only slightly higher and not significantly different from the benchmark. The 14-day recall yields significantly lower mean consumption and higher inequality (thus much higher poverty estimates as in Table 8). By process of elimination, if one recall module must be chosen, the 7-day full list (module 1) presents the fewest complications. Nevertheless, researchers need be aware of the underlying fact that this module is still likely subject to recall and telescoping errors of varying degrees as well as presumably the inability to capture personal out-of-household consumption.[25] For example, the 7-day recall reports significantly less consumption for households with more adult members than does the benchmark, and has a slight tendency to identify larger households as poor. Like other recall modules, it appears subject to either net telescoping or deliberate misreporting that increases in magnitude with the asset wealth of the respondent (Table 6). Nevertheless, this module, as virtually all other modules, yields household consumption rankings that are remarkably stable with respect to key household characteristics, indicating that analysis that focuses on the determinants of household ranking within the consumption distribution will be largely consistent regardless of which module is used. There are long-held perceptions of the inadequacy of using a recall module compared with a household diary to measure consumption expenditures, despite the mixed evidence discussed in Section 2. Our findings call into question these perceptions, particularly when resource implications are considered.

[25] To assess the relative importance of these recall and telescoping errors, a future experiment could compare bounded and unbounded 7-day recall with a suitable benchmark module.

As we have said throughout, the conclusions drawn from this study are dependent on the specific setting. The results presented are likely a result of traits such as the share of consumption from home production, dietary diversity, and the fraction of meals eaten in the dwelling (as opposed to private food consumption at restaurants). These traits observed in Tanzania are mostly associated with low-income countries and not middle-income settings. Rural Tanzania is similar to much of rural Africa and indeed other rural regions in South and South East Asia, and so the general lessons from this study may be broadly applicable to these areas. Urban Tanzania is also not appreciably different from other urban settings in low-income African countries, but may diverge from urban middle-income settings in terms of diversity of consumption choices and household structure. This setting may also diverge from results for parts of West Africa where, as noted by Boozer, Goldstein, and Suri (2009), wives and husbands manage separate budgets for some expenditure items, described as "a decentralized structure," resulting in a higher level of incomplete information about expenditure than observed in the household modules in Tanzania. Further work will be necessary, including the implementation of similar experiments in other settings, to better understand the relative performance of differing approaches to consumption measurement as well as their analytic implications as developing countries continue to raise living standards, increase consumption diversity, and urbanize.

References

Ahmed, A., M. Brzozowski, and T. Crossley. (2006). 'Measurement Errors in Recall Food Consumption Data.' *Working Paper* 06/21 Institute for Fiscal Studies.

Atkinson, A.B., and A. Brandolini. (2001) 'Promise and Pitfalls in the use of "Secondary" Data-Sets: income Inequality in OECD Countries as a Case Study.' *Journal of Economic Literature* 39(3): 771-799.

Bardasi, E., K. Beegle, A. Dillon, and P. Serneels. (2010). 'Do Labor Statistics Depend on How and to Whom the Questions are Asked? Results from a Survey Experiment in Tanzania.' World Bank Policy Research Working Paper 5192.

Bhalla, S. (2000). 'Growth and Poverty in India – Myth and Reality.' in Govinda Rao (ed.), *Poverty and Public Policy: Essays in Honour of Raja Chelliah*, Oxford University Press.

Boozer, M.A., M. Goldstein, and T. Suri. (2009). 'Household Information: Implications for Poverty Measurement and Dynamics.' Mimeo.

Capéau, B., and S. Dercon. (2006).'Prices, Unit Values and Local Measurement Units in Rural Surveys: an Econometric Approach with an Application to Poverty Measurement in Ethiopia.' *Journal of African Economies* 15(2): 181-211.

Chen, S., and M. Ravallion. (2010). 'The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight against Poverty.' *Quarterly Journal of Economics* 125(4): 1577-1625.

Comerford, D., L. Delaney, and C. Harmon. (2009).'Experimental Tests of Survey Responses to Expenditure Questions.' IZA Discussion Paper No. 4389.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins, Baltimore.

Deaton, A. (2005). 'Measuring Poverty in a Growing World (or Measuring Growth in a Poor World).' *Review of Economics and Statistics* 87(1): 1-19.

Deaton, A., and M. Grosh. (2000). 'Consumption.' in Grosh, Margaret and Paul Glewwe, eds. Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study. Washington, D.C.: World Bank.

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

Deaton, A., and S. Zaidi. (2002). 'Guidelines for Constructing Consumption Aggregates for Welfare Analysis' World Bank LSMS Working Paper 135.

Deininger, K., and L. Squire. (1996). 'A New Data Set Measuring Income Inequality.' *The World Bank Economic Review* 10(3): 565-591.

Dutta Roy, D.K., and S.J. Mabey. (1968). 'Household Budget Survey in Ghana.' Technical Publication Series No. 2. Institute of Statistics, University of Ghana, Legon.

Filmer, D., and L.H. Pritchett. (2001). 'Estimating Wealth Effect Without Expenditure Data or Tears: An Application to Educational Enrollments in States of India.' *Demography* 38(1): 115-32.

Freedman, D.A. (2008). 'On Regression Adjustments to Experimental Data.' *Advances in Applied Mathematics* 40(2): 180-193.

Gibson, J. (1999). 'Are Poverty Comparisons Robust to Changes in Household Survey Methods?' Mimeo.

Gibson, J. (2006). 'Statistical Tools and Estimation Methods for Poverty Measures Based on Cross-Sectional Household Surveys.' in *Handbook on Poverty Statistics*, ed. Jonathan Morduch. United Nations Statistics Division.

Gibson, J. (2007) 'A guide to using prices in poverty analysis' http://siteresources.worldbank.org/INTPA/Resources/429966-1092778639630/Prices_in_Poverty_Analysis_FINAL.pdf The World Bank.

Gibson, J., and A.A. Poduzov. (2003). 'Assessing Welfare Indicators for Poverty Measurement in Russia.' Mimeo.

Gibson, J., S. Rozelle, and J. Huang. (2003). 'Improving estimates of inequality and poverty from urban China's Household Income and Expenditure survey.' *Review of Income and Wealth* 49(1): 53-68.

Gieseman, R. (1987). 'The Consumer Expenditure Survey: Quality Control by Comparative Analysis.' *Monthly Labor Review* 110(8): 8-14.

Grosh, M., Q. Zhao, and H. Jeancard. (1995). 'The Sensitivity of Consumption Aggregates to Questionnaire Formulation: Some Preliminary Evidence from the Jamaican and Ghanaian LSMS Surveys.' World Bank, Policy Research Department mimeo.

Jolliffe, D. (2001). 'Measuring Absolute and Relative Poverty: The Sensitivity of Estimated Household Consumption to Survey Design.' *Journal of Economic and Social Measurement* 27(1): 1-23.

Kulshrestha, A., and A. Kar. (2005). 'Consumer Expenditure from the National Accounts and National Sample Survey.' In A. Deaton and V. Kozel (eds.), *The Great Indian Poverty Debate* New Delhi, India: MacMillan, Chapter 7.

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

Lanjouw, P. (2005). 'Constructing a Consumption Aggregate for the Purpose of Welfare Analysis: Issues and Recommendations Concerning the POF 2002/3 in Brazil.' Development Economics Research Group, World Bank. Mimeo.

Lanjouw, J., and P. Lanjouw. (2001). 'How to Compare Apples and Oranges: Poverty Measurement Based on Different Definitions of Consumption.' *Review of Income and Wealth* 47(1):25-42.

McWhinney, I., and H. Champion. (1974). 'The Canadian Experience with Recall and Diary Methods in Consumer Expenditure Surveys.' *Annals of Economic and Social Measurement* 3(2): 411-435.

Minhas, B. (1988). 'Validation of Large-scale Sample Survey Data: Case of NSS Household Consumption Expenditure.' *Sankhya Series B* 50(3): S1-S63.

NBS (National Bureau of Statistics) (2002). 'Household Budget Survey 2000/01.' NBS: Dar es Salaam.

National Economic Council. (2000). 'Profile of Poverty of Malawi, 1998: Poverty Analysis of the Malawi Integrated Household Survey, 1997-98.' Mimeo.

Neter, J. (1970). 'Measurement Errors in Reports of Consumer Expenditures.' *Journal of Marketing Research* 7: 11-25.

Neter, J. and J. Waksburg. (1964). 'A Study of Response Errors in Expenditure Data From Household Interviews.' *Journal of the American Statistical Association* 59(305): 18-55.

NSSO Expert Group on Sampling Errors. (2003). 'Suitability of Different Reference Periods for Measuring Household Consumption: Results of a Pilot Study.' *Economic and Political Weekly*, 37(4): 307–321.

Pinkovskiy, M., and X. Sala-i-Martin. (2009). 'Parametric Estimations of the World Distribution of Income.' NBER Working Paper 15433.

Plewis, I. (1972). 'Experiments on Expenditure and Recall Periods, June-September 1970.' Statistical Newsletter No. 40. U.N. Economic Commission for Africa, Addis Ababa, Ethiopia.

Pradhan, M. (2001). 'Welfare Analysis with a Proxy Consumption Measure: Evidence from a Repeated Experiment in Indonesia.' Mimeo.

Ravallion, M. (1996). "Issues in Measuring and Modelling Poverty". *The Economic Journal* 106(5): 1328-1343.

Ravallion, M. (2003). 'Measuring Aggregate Welfare in Developing Countries: How Well Do National Accounts and Surveys Agree?' *Review of Economics and Statistics* 85(3): 645–652.

Sala-i-Martin, X., and M. Pinkovskiy. (2010). 'African Poverty is Falling…Much Faster than You Think!' NBER Working Paper 15775.

Scott, C., and B. Amenuvegbe. (1991). 'Recall Loss and Recall Duration: an Experimental Study in Ghana.' *Inter-Stat* 4(1): 31-55.

Statistical Institute and Planning Institute of Jamaica. (1994). 1994 Survey of Living Conditions. Kingston.

Steele, D. (1998). 'Ecuador Consumption Items.' World Bank. Mimeo.

United Nations Statistics Division (2005). *Handbook on Poverty Statistics: Concepts, Methods and Policy Use* United Nations Department of Economic and Social Affairs, Etudes Méthodologiques (Series F), No.99.

World Bank. (1993). 'Indonesia: Public Expenditures, Prices and the Poor.' Report 11293-IND, Indonesia Resident Mission, Jakarta.

World Bank. (2005). 'Russian Federation: Reducing Poverty through Growth and Social Policy Reform.' Report 28923-RU, World Bank.

World Bank. (2006). 'Bosnia and Herzegovina: HBS-LSMS Comparison.' Eastern Europe and Central Asia Poverty Reduction and Economic Management Unit. Mimeo.

Yu, D. (2008). 'The Comparability of Income and Expenditure Surveys 1995, 2000, and 2005/2006.' Steenbosch Economic Working Papers 11/08.

Table 1. Survey experiment consumption modules

Module	Consumption measurement	Food content	N households
1	Long list (58 food items) 14 day	Quantity from purchases, own-production, and gifts/other sources; Tshilling value of consumption from purchases	504
2	Long list (58 food items) 7 day	Quantity from purchases, own-production, and gifts/other sources; Tshilling value of consumption from purchases	504
3	Subset list (17 food items; subset of 58 foods) 7 day	Quantity from purchases, own-production, and gifts/other sources; Tshilling value of consumption from purchases	504
4	Collapsed list (11 food items covering universe of food categories) 7 day	Tshilling value of consumption	504
5	Long list (58 food items) Usual 12 month	Consumption from purchases: number of months consumed, quantity per month, Tshilling value per month Consumption from own-production: number of months consumed, quantity per month, Tshilling value per month Consumption from gifts/other sources: total estimated value for last 12 months	504
6	Household diary, frequent visits 14 day diary		503
7	Household diary, infrequent visits 14 day diary		503
8	Personal diary, frequent visits 14 day diary		503
			4,029

Notes: Frequent visits entailed daily visits by the local assistant and visits every other day by the survey enumerator for the duration of the 2-week diary. Infrequent visits entail 3 visits: to deliver the diary (day 1), to pick up week 1 diary and drop off week 2 diary (day 8), and to pick up week 2 diary (day 15). Households assigned to the infrequent diary but who had no literate members (about 18 percent of the 503 households) were visited every other day by the local assistant and the enumerator.

Non-food items are divided into two groups based on frequency of purchase. Frequently purchased items (charcoal, firewood, kerosene/paraffin, matches, candles, lighters, laundry soap, toilet soap, cigarettes, tobacco, cell phone and internet, transport) were collected by 14-day recall for modules 1-5 and in the 14-day diary for modules 6-8. Non-frequent non-food items (utilities, durables, clothing, health, education, contributions, and other; housing is excluded) are collected by recall identically across all modules at the end of the interview (and at the end of the 2-week period for the diaries) and over the identical one or 12-month reference period, depending on the item in question.

Table 2. Consumption expenditure per capita (annualized Tanzania shillings) by consumption module

	Mean				Median			
Module	Total	Food	Non-food frequent (recall or diary)	Non-food non-frequent (all recall)	Total	Food	Non-food frequent (recall or diary)	Non-food non-frequent (all recall)
1. Long 14 day	485,251	330,281	66,875	88,095	268,434	210,059	17,064	37,375
2. Long 7 day	520,850	356,571	70,371	93,908	311,955	245,045	18,752	39.840
3. Subset 7 day	489,637	321,432	73,022	95,183	304,473	237,602	20,306	39,560
4. Collapse 7 day	408,783	257,125	67,136	84,522	239,503	177,602	18,249	36,605
5. Long usual 12 month	486,181	294,505	88,037	103,638	254,735	182,769	21,377	41,679
6. HH diary frequent	412,843	271,038	52,372	89,433	279,779	210,192	15,870	38,525
7. HH diary infrequent	425,298	292,511	48,685	84,102	290,870	218,647	17,147	39,033
8. Personal diary	510,616	347,671	68,556	94,389	329,847	249,083	20,322	42,425
All modules	467,840	308,913	66,902	91,665	287,732	216,169	18,523	39,635

Table 3. Regressions of per capita consumption expenditure, by total, food, and non-food

	(1)	(2)	(3)	(4)
			ln non-food,	ln non-food,
			frequent	non-frequent
Personal diary omitted	Ln total	ln food	(recall or diary)	(all recall)
1. Recall: Long, 14 day	-0.161***	-0.167***	-0.104	-0.105*
	(0.037)	(0.037)	(0.067)	(0.060)
2. Recall: Long, 7 day	-0.039	-0.017	-0.134**	-0.096
	(0.037)	(0.037)	(0.067)	(0.060)
3. Recall: Subset, 7 day	-0.071*	-0.079**	-0.112*	-0.090
	(0.037)	(0.037)	(0.067)	(0.060)
4. Recall: Collapse, 7 day	-0.283***	-0.332***	-0.104	-0.138**
	(0.037)	(0.037)	(0.067)	(0.060)
5. Recall: Long usual 12 month	-0.207***	-0.268***	0.023	-0.013
	(0.037)	(0.037)	(0.067)	(0.060)
6. Diary: HH, frequent	-0.173***	-0.196***	-0.279***	-0.046
	(0.037)	(0.037)	(0.067)	(0.060)
7. Diary: HH, infrequent	-0.136***	-0.129***	-0.244***	-0.105*
	(0.037)	(0.037)	(0.067)	(0.060)
Number of households	4,025	4,025	3,942	4,016

Note: *** indicates significance at 1 percent; ** at 5 percent; and * at 10 percent. Standard errors in parentheses. Columns 3 and 4 have smaller sample sizes due to some households with zero non-food expenditures. 83 households have no frequent non-food expenditures (6, 6, 1, 9, 7, 17, 17, and 20 households for modules 1-8, respectively). 9 households have no non-frequent non-food expenditures (3, 3, 1, and 2 households for modules 5, 6, 7 and 8, respectively).

Table 4. Comparison of recall modules (per capita consumption)

	(1)	(2)	(3)	(4)
Long, 7 day omitted	Ln total	ln food	ln non-food, frequent (recall or diary)	ln non-food, non-frequent (all recall)
Panel A.				
1.Long, 14 day	-0.121***	-0.151***	0.032	-0.008
	(0.037)	(0.037)	(0.062)	(0.059)
5.Usual 12 month	-0.168***	-0.251***	0.158**	0.084
	(0.037)	(0.037)	(0.062)	(0.059)
Number of households	1,511	1,511	1,492	1,508
Panel B.				
3.Subset, 7 day	-0.032	-0.063*	0.021	0.006
	(0.036)	(0.036)	(0.063)	(0.061)
4.Collapse, 7 day	-0.244***	-0.315***	0.027	-0.042
	(0.036)	(0.036)	(0.063)	(0.061)
Number of households	1,512	1,512	1,496	1,512

Note: *** indicates significance at 1 percent; ** at 5 percent; and * at 10 percent. Standard errors in parentheses. Columns 3, 4, 7 and 8 have smaller sample sizes due to some households with zero non-food expenditures. Excluded category in both regressions is module 2 – the full-list 7-day recall.

Table 5. Comparison of diaries (per capita consumption)

	(1)	(2)	(3)	(4)
			ln non-food,	ln non-food,
			frequent	non-frequent
Personal diary omitted	Ln total	ln food	(recall or diary)	(all recall)
Panel A.				
6. Diary: HH, frequent	-0.173***	-0.195***	-0.277***	-0.045
	(0.036)	(0.035)	(0.069)	(0.059)
7. Diary: HH, infrequent	-0.135***	-0.129***	-0.245***	-0.103*
	(0.036)	(0.035)	(0.069)	(0.059)
Number of households	1,506	1,506	1,452	1,499
Panel B.				
6. Diary: HH, frequent	-0.185***	-0.221***	-0.259***	0.021
literate	(0.037)	(0.037)	(0.072)	(0.060)
6. Diary: HH, frequent	-0.099	-0.038	-0.384***	-0.467***
illiterate	(0.074)	(0.073)	(0.146)	(0.122)
7. Diary: HH, infrequent	-0.109***	-0.112***	-0.155**	-0.024
literate	(0.038)	(0.037)	(0.073)	(0.061)
7. Diary: HH, infrequent	-0.252***	-0.202***	-0.686***	-0.514***
illiterate	(0.069)	(0.068)	(0.136)	(0.112)
Number of households	1,506	1,506	1,452	1,499

Note: *** indicates significance at 1 percent; ** at 5 percent; and * at 10 percent. Standard errors in parentheses. Columns 3, 4, 7, and 8 have smaller sample sizes due to some households with zero non-food expenditures. Excluded category in both panels is module 8 – frequently supervised individual diary.

Households assigned to a household diary with infrequent visits were visited slightly more often if illiterate than literate (see Table 1 and text for further explanation).

Table 6. Interaction of consumption module and select household characteristics

	Panel A *Interactions with log per capita total consumption*					**Panel B** *Interactions with log per capita food consumption*				
Personal diary omitted	Household size	Number of adults	Urban	Education of hh head	Asset index	Household size	Number of adults	Urban	Education of hh head	Asset index
1. Recall: Long, 14 day	-0.047***	-0.123***	0.112	-0.006	0.075**	-0.046***	-0.113***	0.140	0.001	0.095**
2. Recall: Long, 7 day	-0.028	-0.101**	0.092	-0.001	0.081**	-0.023	-0.078**	0.144	0.006	0.108***
3. Recall: Subset, 7 day	-0.039**	-0.078*	0.092	0.004	0.086**	-0.032*	-0.059	0.079	0.003	0.074*
4. Recall: Collapse, 7 day	-0.025	-0.060	-0.004	0.005	0.086**	-0.017	-0.044	-0.018	0.010	0.080**
5. Recall: Long, usual month	-0.042**	-0.086**	0.175*	0.025**	0.149***	-0.037**	-0.072*	0.203**	0.028**	0.160***
6. Diary: HH, frequent	-0.009	-0.042	-0.115	-0.006	-0.028	-0.003	-0.035	-0.103	-0.007	-0.038
7. Diary: HH, infrequent	0.013	-0.038	-0.286***	-0.005	-0.072*	0.014	-0.016	-0.256***	-0.001	-0.055
	Panel C *Interactions with Log per capita frequent non-food consumption*					**Panel D** *Interactions with log per capita non-frequent non-food consumption*				
1. Recall: Long, 14 day	-0.078**	-0.217***	0.107	-0.006	0.087	-0.031	-0.121*	-0.053	-0.033*	-0.024
2. Recall: Long, 7 day	-0.094***	-0.218***	0.150	0.002	0.151**	-0.014	-0.131**	-0.054	-0.013	0.026
3. Recall: Subset, 7 day	-0.066*	-0.131	0.234	0.025	0.180***	-0.034	-0.078	0.033	0.005	0.044
4. Recall: Collapse, 7 day	-0.076**	-0.132*	-0.013	0.002	0.148**	-0.020	-0.068	-0.084	-0.024	-0.015
5. Recall: Long, usual month	-0.102***	-0.187**	0.194	0.022	0.143**	-0.057*	-0.149**	-0.038	-0.002	0.017
6. Diary: HH, frequent	0.002	-0.075	-0.301*	-0.012	-0.041	-0.012	-0.050	-0.295**	-0.015	-0.047
7. Diary: HH, infrequent	0.020	-0.107	-0.452***	-0.022	-0.154**	0.009	-0.073	-0.311**	-0.015	-0.065

Note: estimates of Equation (2). Each column represents the results of a (separate) OLS of a selected consumption expenditure measure (mentioned in the titles of the panels) on 7 module assignment dummies, a single selected household characteristic (mentioned in the column headings) and 7 interaction terms of that household characteristic with the module assignment dummies. Only the interaction terms are reported. The personal diary is the omitted category. Level effects and standard errors are omitted to improve readability, but available upon request from the authors. *** indicates significance at 1 percent; ** at 5 percent; and * at 10 percent.

Table 7. Gini coefficient by survey module

	Total per capita	Food consumption per capita	Frequent non-food per capita	Non-frequent non-food per capita
1. Long 14 day	0.512***	0.477***	0.713	0.625
	(0.019)	(0.020)	(0.017)	(0.020)
2. Long 7 day	0.483	0.432	0.730	0.632
	(0.020)	(0.017)	(0.027)	(0.018)
3. Subset 7 day	0.467	0.403	0.725	0.641
	(0.014)	(0.012)	(0.015)	(0.020)
4. Collapse 7 day	0.484	0.424	0.712	0.622
	(0.021)	(0.018)	(0.017)	(0.026)
5. Long Usual 12 month	0.537***	0.470***	0.746	0.653*
	(0.019)	(0.015)	(0.030)	(0.021)
6. HH diary Frequent	0.427	0.361**	0.702	0.609
	(0.016)	(0.014)	(0.021)	(0.018)
7. HH diary Infrequent	0.419*	0.372*	0.676	0.607
	(0.017)	(0.016)	(0.024)	(0.020)
8. Personal diary	0.450	0.403	0.706	0.610
	(0.017)	(0.015)	(0.023)	(0.017)

Note: Jackknife standard errors in parentheses. *** indicates significant difference compared with module 8 (personal diary) at 1 percent; ** at 5 percent; and * at 10 percent.

Table 8. Poverty statistics by survey module

	Poverty line at $1.25/person/day			Poverty line at $0.78/person/day
	Poverty headcount	Poverty gap ratio	Squared gap (x100)	Poverty headcount
1. Long 14 day	62.8**	25.8***	13.4***	34.9***
	(2.9)	(1.7)	(1.1)	(3.0)
2. Long 7 day	54.9	19.1	9.0	22.9
	(3.0)	(1.5)	(1.0)	(2.8)
3. Subset 7 day	55.1	21.4*	10.4*	28.7*
	(3.0)	(1.6)	(1.0)	(2.8)
4. Collapse 7 day	66.8***	28.8***	15.1***	41.1***
	(2.8)	(1.6)	(1.1)	(3.0)
5. Long Usual 12 month	64.6***	28.1***	15.0***	39.4***
	(3.0)	(1.7)	(1.1)	(3.0)
6. HH diary Frequent	59.5**	21.5**	10.2*	28.4*
	(2.9)	(1.4)	(0.9)	(2.6)
7. HH diary Infrequent	55.6	18.9	8.7	22.4
	(2.9)	(1.4)	(0.8)	(2.4)
8. Personal diary	47.5	16.0	7.4	19.8
	(3.1)	(1.3)	(0.7)	(2.4)

Note: Robust standard errors in parentheses. *** indicates significant difference compared with module 8 (personal diary) at 1 percent; ** at 5 percent; and * at 10 percent.

Table 9. Interaction of consumption module and household characteristics with respect to poverty indicator or rank order of household

	Panel A *Poverty indicator($1.25/person/day)*					**Panel B** *Rank order of household (percentile)*				
Personal diary omitted	Household size	Number of adults	Urban	Education of hh head	Asset index	Household size	Number of adults	Urban	Education of hh head	Asset index
1. Recall: Long, 14 day	0.004	0.010	-0.048	0.006	-0.042	-0.878	-3.175**	0.190	-0.486	0.529
2. Recall: Long, 7 day	0.020**	0.023	-0.066	0.004	-0.042	-0.929	-3.344**	1.285	-0.272	1.491
3. Recall: Subset, 7 day	0.007	-0.006	-0.054	-0.000	-0.043	-1.170*	-2.527	1.089	0.052	1.903
4. Recall: Collapse, 7 day	0.005	-0.007	-0.012	-0.005	-0.044	-0.610	-2.131	-3.238	0.064	1.117
5. Recall: Long, Usual month	0.005	-0.023	-0.070	-0.006	-0.060**	-0.716	-2.222	0.178	0.405	1.619
6. Diary: HH, Frequent	-0.021**	-0.033	-0.016	-0.009	-0.030	0.058	-1.316	-3.005	0.010	-0.001
7. Diary: HH, Infrequent	-0.011	0.006	0.025	-0.004	-0.026	0.466	-1.448	-8.783**	0.089	-1.333

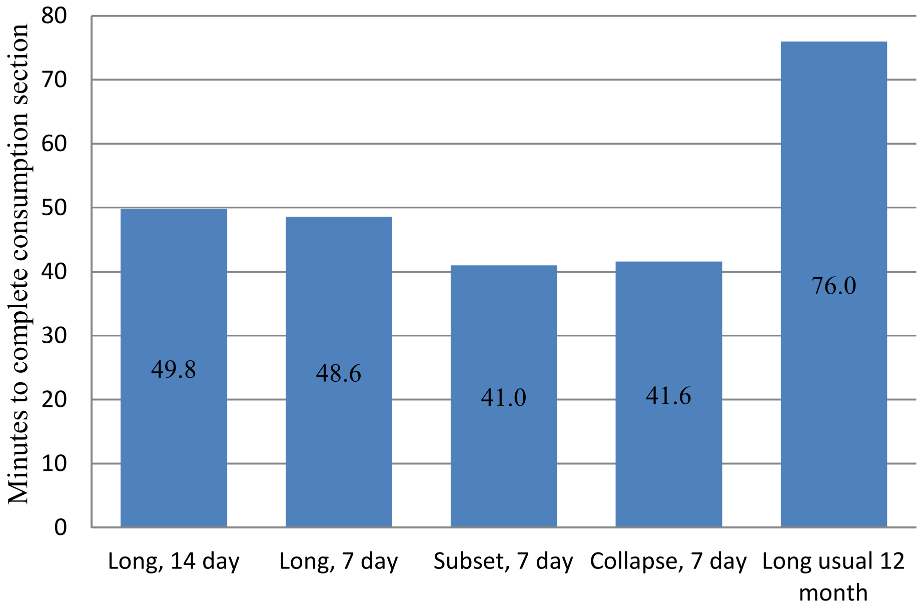
Note: estimates of Equation (2). Each column represents the results of a (separate) OLS of a selected consumption expenditure measure (mentioned in the titles of the panels) on 7 module assignment dummies, a single selected household characteristic (mentioned in the column headings) and 7 interaction terms of that household characteristic with the module assignment dummies. The personal diary is the omitted category. Level effects and standard errors are omitted to improve readability, but available upon request from the authors. *** indicates significance at 1 percent; ** at 5 percent; and * at 10 percent.

Table 10: Cost comparison of survey modules

		(1)	(2)	(3)	(4)	(5)	(6)
Type	Interviewer's visiting schedule per cycle	Total field days	Total HHs completed	HHs interviewed per interviewer per day (fieldwork)	HHs entered per data entry clerk per day (data entry)	Person days time to complete one household (interview and data entry)	Variable cost (US$)
HH Diary Infrequent	weekly visits in 3 EAs with breaks	22	24	1.09	4.31	1.15	334
	weekly visits in 4 EAs no breaks	24	32	1.33	4.31	0.98	280
HH Diary Frequent	daily visits in 1 EA	17	8	0.47	4.31	2.36	726
	visits every 2 days in 2 EAs	20	16	0.80	4.31	1.48	442
Personal Frequent	daily visits in 1 EA	17	6	0.35	2.86	3.18	974
	visits every 2 days in 2 EAs	20	12	0.60	2.86	2.02	597
Recall questionnaire	4 household interviews per day in 1 EA			4.00	8.51	0.37	100 (numeraire)

Note: EA is enumeration area or primary sampling unit (i.e., village/community). Field days are twice as expensive as data entry days. Calculations assume no substantial difference in daily transport costs across survey options. Time to interview includes arrive, diary introductions to households, and household basic information questionnaire completion. Column 5 is [1/(column 3) + 1/(column 4)].

Figure 1. Mean time of completion for recall modules



Appendix Table 1. Basic household characteristics by consumption module assignment

	Recall Module					Diary Module		
	1	2	3	4	5	6	7	8
Head: female	18.8	19.8	20.6	20.8	21.6	18.1	21.1	20.9
Head: age	47.6	46.2	46.0	46.4	46.5	46.6	47.0	46.8
Head: years of schooling	4.7	4.8	4.7	4.7	4.8	4.7	4.7	4.7
Head: married	74.2	73.2	72.0	73.0	71.8	74.8	73.8	76.3
Household size*	5.2	5.2	5.2	5.5	5.3	5.3	5.3	5.3
Adult equivalent household size*	4.1	4.1	4.1	4.3	4.2	4.2	4.2	4.3
Share of members less 6 years	17.6	17.0	18.3	18.7	17.7	17.3	17.7	17.7
Share of members 6-15 years	24.1	24.8	24.8	25.0	24.8	24.3	23.5	24.3
Concrete/tile flooring (non-earth)	25.8	25.4	26.0	25.8	26.4	25.8	26.2	25.0
Main source for lighting is electricity/generator/solar panels	11.7	10.7	9.3	11.1	11.7	11.7	10.7	10.7
Owns a mobile telephone*	30.8	30.6	29.4	29.8	30.8	34.0	29.8	28.8
Bicycle*	43.1	44.2	40.5	44.2	42.3	43.7	48.1	46.5
Owns any land	80.6	78.4	78.4	79.0	81.3	80.9	79.3	81.9
Acres of land owned (incld 0s)	3.3	3.1	3.1	3.5	3.5	3.2	3.5	3.2
Month of interview (1=Jan,12=Dec)	5.9	5.9	6.0	6.0	6.0	5.8	5.8	5.8
Number of households	504	504	504	504	504	503	503	503

Notes: * indicates statistical difference in mean across at least two pairs at 5 percent. See NBS (2002) for details on the adult equivalence scales.