2019

IARIW-World Bank

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

Effects of Data Collection Methods on Estimated Household Consumption and Poverty, and on Survey Costs: Evidence from an Experiment in the Marshall Islands

Michael Sharp

Bertrand Buffière

Kristen Himelein

John Gibson

Paper Prepared for the IARIW-World Bank Conference

Washington, DC, November 7-8, 2019

Session 1A: Methods for Measuring Consumption and Poverty

Time: 11:00 – 13:00, September 7

October 2019
Preliminary – not to be cited

Effects of Data Collection Methods on Estimated Household Consumption and Poverty, and on Survey Costs: Evidence from an Experiment in the Marshall Islands

Michael Sharp, Bertrand Buffière, Kristen Himelein and John Gibson[1]

Household Income and Expenditure Surveys (HIES) are the main source of household-level consumption and expenditure information in many countries. These data serve many purposes, including poverty and food security monitoring, updating the CPI basket, and as an input to National Account calculations. In the Pacific, these surveys have historically gathered expenditure data with open form diaries completed on paper. This methodology, however, is costly to collect and burdensome for respondents, leading to long lags between rounds, substantial processing time, and decreased policy relevance due to a lack of frequency and timeliness. Also, non-compliance and partial-compliance with the diary-keeping method complicates analyses because it is hard to separate households with low consumption from those who lose enthusiasm for keeping the diary. This paper reports on a survey experiment carried out in the Marshall Islands where two variants of existing 14-day diary-keeping HIES used in the Pacific are compared with a 14-day diary survey that uses computer-assisted personal interviewing (CAPI) and with two variants of 7-day recall surveys using CAPI. Variable costs for the *status quo* diary survey design are between 2.8 and 4.4 times as expensive as for a single-visit 7-day recall survey, while the CAPI diary is even more costly. Despite the high cost of the diary-keeping surveys, they yield data that are, overall, of worse quality. The effective completion rate with less monitored diaries is only two-thirds and apparent consumption is significantly lower (and poverty higher) compared to all other modules. The highly monitored diaries give similar results to using recall, but at much greater cost.

JEL: C81, O15, Q18

Keywords: CAPI, Diary, Experiment, Recall, Survey Design, Hunger, Poverty

Acknowledgements: We are grateful to comments from the Pacific Statistics Methods Board and Kathleen Beegle. All remaining errors are those of the authors.

[1] Sharp and Buffière are with the Statistics for Development Division of the Pacific Community, Himelein is with the Poverty and Equity Global Practice of the World Bank, Gibson is with the Department of Economics, University of Waikato. These are the views of the authors and need not reflect those of their employers.

I. Introduction

In much of the world, data from Household Income and Expenditure Surveys (HIES) are used for multiple objectives, including poverty monitoring, updating the CPI basket, and as an input to National Account calculations. These surveys are very costly to field, especially if using the diary-keeping method. For example, in the Pacific Islands where the current study is set, recent surveys that involved interviewers staying in villages for 3-weeks in order to oversee 14-day diaries had total survey costs per completed household of almost US$700 in the Solomon Islands and around US$4000 per household in Papua New Guinea (PNG). Even in less challenging environments, in the smaller atoll countries, costs are up to $800 per surveyed household. These high costs also mean that HIES are fielded only infrequently, with an average gap of nine years between each HIES in the typical Pacific country. This infrequency limits the usefulness of data from these surveys, especially for poverty monitoring.

Also, diary-keeping surveys are very burdensome for respondents, with considerable *diary fatigue*. The resulting non-compliance and partial-compliance due to fatigue complicates analyses because it is hard to separate households with genuinely low consumption from those who either stop keeping the diary or who revert to an informal recall survey when interviewers revisit. For example, the 2009/10 PNG HIES saw the number of transactions listed in the diaries decline by 3.4 percent per day, on average, over the 14-day reporting period. This fatigue caused a decline in apparent consumption that produces a concomitant rise in apparent poverty; if only the first seven days of diary records are used the headcount poverty rate in PNG in 2009/10 would have been 41%, versus 47% if only the second seven days are used (Gibson, 2013). A related issue with the time-consuming approach to measuring commodity inflows and outflows with diaries is that it limits the scope to focus on other topics; thus, in contrast to multi-topic

surveys (such as the LSMS), fewer correlates of poverty are available in diary-keeping surveys and other key data needed for modelling and interpreting nominal consumption data, such as local prices, are almost entirely absent.

In light of these issues, this paper provides a report on a survey experiment carried out in 2018 in the Republic of the Marshall Islands (RMI). The survey fieldwork was carried out by staff of the Economic Policy, Planning and Statistics Office (EPPSO) of the Government of the Marshall Islands, with design and analysis of the experiment reflecting a collaboration between the Secretariat of the Pacific Community (SPC) and the World Bank. The experiment was designed to provide evidence to aid development of more cost-effective and reliable survey approaches that will support future poverty monitoring and meet the diverse demands placed on HIES data in the Pacific. Though the recent survey methodology literature from both developed and developing countries has generally concluded that recall data collection may be more effective than traditional diaries (Zezza et al, 2017), especially when survey resources are constrained, it is not clear how these results translate to Pacific Island countries, where much consumption is characterized by bulk purchases of imported goods and continuous harvesting of fish and tubers. Our experiment is also forward-looking by studying the performance of these data-collection methods within a computer assisted personal interview (CAPI) format.

Specifically, the experiment had five arms, each fielded within the same enumeration areas at the same time:

1. 14-day diary, highly monitored (with interviewer visits every two days), with transactions recorded with pen and paper (PAPI), and coding and data entry by EPPSO after the field work was completed
2. 14-day diary, less monitored (with interviewer visits after each week), with transactions recorded with pen and paper (PAPI), and coding and data entry by EPPSO after the field work was completed
3. 14-day diary, highly monitored (with interviewer visits every two days), with transactions transcribed into the tablets using CAPI during each interviewer visit

4. 7-day single-visit unbounded recall, using a list of 102 food groups and 20 non-food groups, with data entered into the tablets using CAPI during the interview
5. 7-day two-visit recall, with an initial visit made to the household to indicate the start of the recall period (and to gather other data), using a recall list of 102 food groups and 20 non-food groups, with data entered into the tablets using CAPI during the interview.[2]

Some combination of arms 1 and 2 reflects the *status quo* for HIES in the Pacific; statistics offices may intend to use highly monitored diaries but without strict supervision and a generous budget for labour and travel for interviewer revisits, the survey can degrade into less monitored diaries (that may further devolve into a pseudo-recall survey). In addition to the basic difference in data reporting method, of diary versus recall, our experimental arms also differ in whether they directly ask about consumption (arms 4 and 5) or instead indirectly derive consumption as a residual from the following components:

	Purchases
+	Own-production
+	Gifts received
-	Sales
-	Gifts given
-	Net stock increases
=	Consumption.

A key aspect of the indirect approach is that arms 1, 2, and 3 attempt to measure starting and ending food stocks, because otherwise the acquisitions-based diaries may understate the food available (and the value of food consumption) if the household consumes from existing stocks, or overstate if acquisitions during the diary-keeping period go into ending stocks. This intrusive approach is somewhat idiosyncratic to the Pacific, which features non-seasonal agriculture, bulky root crops, and high transactions costs of going to markets and gardens that results in

[2] However, the recall questions continued to use the "In the last 7 days…" format and the gap between visits varied. For example, visit 2 was 8 days after visit 1 for 53% of households, 9 or more days after for 14% of households, and 7 or less days after for 33% of households. Thus, this experiment arm does not properly test bounded recall.

considerable in-home storage and hidden consumption from stocks (Gibson and Kim, 2012).

In our experiment the *status quo* diary-keeping surveys have total costs that range from US$1160 to US$3020 per household in the rural sector (and $550 to $1220 in urban areas). The range reflects whether revisit frequency was every two days or every week, which affects labour costs and travel costs. The costs were even higher with the CAPI diary, because of interviewer time to transcribe transactions from paper diaries into tablets. In contrast, single-visit recall has a total cost of US$580 per household in the rural sector and $290 in the urban sector. These high costs partly reflect issues with surveying in small, scattered, atolls and so we also have variable cost estimates that should be more transferrable across settings; these suggest *status quo* diary surveys are between 2.8 and 4.4 times as expensive as for a single-visit 7-day recall survey, and if these diaries used CAPI (with interviewer transcription) the cost would be five-times higher. Although the setting is very different, these cost ratios are close to what Beegle et al (2012) estimate for Tanzania, that a household diary with interviewer visits every two days for 14 days is 4.4 times as expensive as a single-visit recall survey and if the diary-checking visits are only once per week the diary is from 2.8 to 3.3 times as expensive as the single visit recall survey.

Despite the extra cost of the diary-keeping surveys, they yield data that are, overall, of worse quality. The effective completion rate with less monitored diaries is only two-thirds and apparent consumption is significantly lower (and poverty higher) compared to all other modules. The highly monitored diaries give similar results to using recall, but at much greater cost.

The remainder of the paper is as follows: Section 2 briefly reviews related literature and previous evidence from the Pacific, Section 3 describes the experiment and survey costs, while results for consumption and poverty measurement are in Section 4. Section 5 provides discussion as well as recommendations for future data collection and areas of further research.

2. Literature Review

Theoretically a well-implemented diary should provide more accurate results and better granularity than recall data. Recall data asks respondents to remember while diaries are in principle collected in real time and therefore not subject to recall error. Recall error stems from three main sources: heaping, telescoping, and omission. Heaping describes the rounding of values to even numbers, such as estimating total purchases as 5 kg instead of 4.8 kg or 5.2 kg. While our experimental data are likely subject to some degree of heaping, we will not examine it in detail here as market purchases are often naturally heaped: $2 worth of coconuts or 1 kg of rice. In addition, all arms would be subject to approximately the same level of heaping. Telescoping describes inaccurately dating relevant actions, either placing distant events more recently (forward telescoping) or pushing recent events further back in time (backwards telescoping). Both types of error can impact the expenditure measure since items would be inaccurately included or excluded from the recall period. To evaluate the potential impact of telescoping, our experiment was intended to include both bounded and unbounded recall periods because bounding has been found to decrease telescoping and increase accuracy (Loftus and Marburger, 1983). However, the way the two-visit recall was implemented precludes a proper test of bounding effects.

The third source of error, omission, refers to excluding relevant events which took place. Omission error in expenditure data could occur because respondents forget certain transactions when asked to remember over the recall period. For example, Scott and Amenuvegbe (1991) find that average daily expenditures reported by survey respondents in Ghana fell by almost three percent for every day added to the recall period, with the greatest decline for the more frequently purchased items; this pattern is dubbed “progressive amnesia” by Deaton (1997).

However, omission could also occur with diary data collection, if respondents do not enter all relevant transactions into the diary, due to either forgetting or because they find the burden of compliance too high. Relatedly, Schündeln (2018) finds that a succession of up to 10 visits to a household within a month, to create a monthly measure of consumption as the sum of a series of 3-day recalls, leads to monotonically declining compliance with each successive visit. This pattern of concentrated revisits is similar to what occurs with a highly monitored diary, and so a pattern of declining compliance with diary surveys may also occur.

Whether omission error is higher in recall or diary data collection and how supervision affects omission error are central questions for making cost-effective methodology decisions. In their review of the literature, Eisenhower et al. (1991) concluded that short recall periods were more impacted by telescoping than by recall error, but that telescoping was more likely for larger expenses while smaller items, in particular food and other routine purchases, are more likely to be omitted. Likewise, Friedman et al. (2017) found that a 7-day food recall overstated the value of consumption, conditional on incidence, by more than for a 14-day recall, which is consistent with forward telescoping and the misdated consumption being amortized over a longer period (and, hence, averaging to a smaller error) with 14-day recall. For both periods, the incidence of consumption for most food groups was understated, compared to the benchmark, and in their setting (Tanzania) incidence errors and value errors approximately cancelled out for the 7-day recall survey (while incidence errors outweighed value errors for the 14-day recall).

Specifically related to food data collection, there is growing evidence to support recall approaches over the diary method. In their introduction to a special issue of *Food Policy* on the current international best practices for food data collection, Zezza et al. (2017) review the literature from both developed and developing countries and conclude, “recall surveys tend to

return higher consumption values (whether in monetary or caloric terms) than diaries." This conclusion draws from evidence ranging from Niger, one of the world's poorest countries where Backiny-Yetna et al. (2017) found per capita consumption from a 7-day diary was 28 percent lower than what was found with a 7-day recall, to Canada, where Brzozowski et al. (2017) found shortcoming in both diary and recall measures.[3] Bee et al. (2017) reviewed recall and diary data collection as background to updating the Consumption Expenditure Survey, which is the main source of microdata on consumption for the United States, and concluded that recall provided higher and more accurate measures of consumption across all major categories, and that "using diary data to assess inequality trends and other distributional outcomes is likely to lead to biased and misleading results."

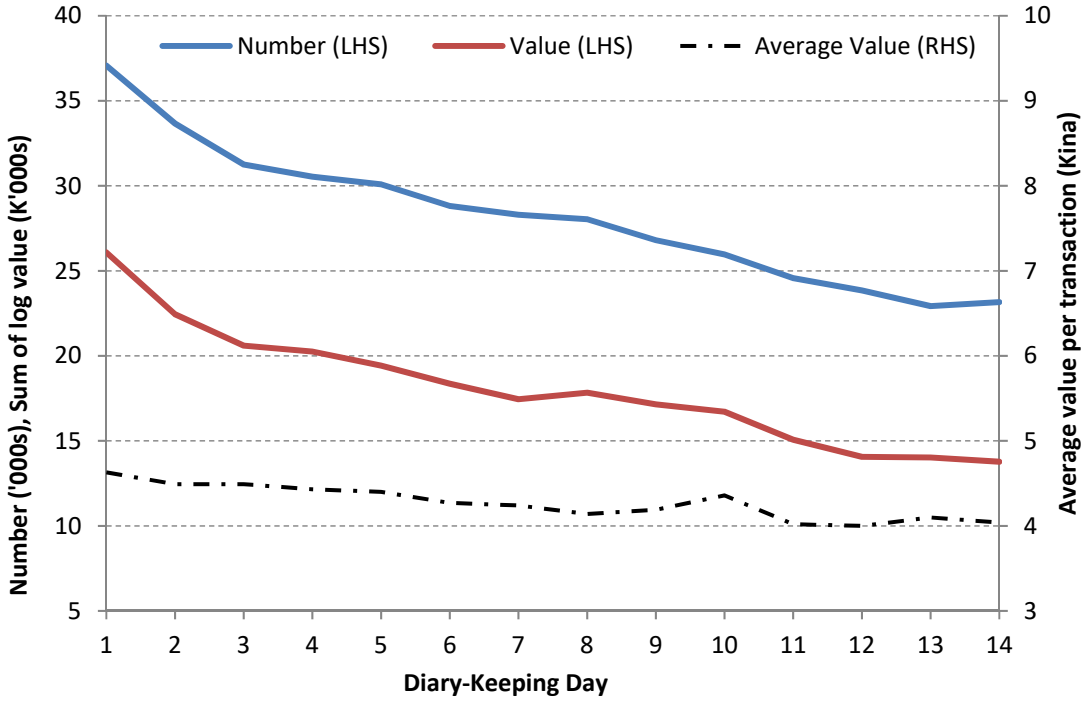
2.1 Prior Evidence from the Pacific

The various HIES carried out in the Pacific prior to the current experiment all use the diary method. As such, evidence from the region mainly deals with the pathologies of diaries. Clear evidence of diary fatigue comes from the 2009/10 PNG HIES, which had a 14-day recording period, with interviewers living in each village for three weeks so that they could check on the diary-keeping households every second day, or so. The diary-keeping was staggered, both over the months of the year, and also the days of the week. Across all of the 3800 households who provided diary data (from a target of 4100), a total of 37,000 transactions were recorded on the first day of diary-keeping (10 per household); these covered all forms of acquisition, such as purchases or self-production. However, by day 14 of diary-keeping, the total transactions were

[3] While Beegle et al. (2012) found a 7-day recall module gave food and total consumption expenditure in Tanzania that most closely matched the gold-standard of a highly supervised individual diary, compared to the performance of six other alternative designs, including frequent and infrequently monitored household level diaries, the subsequent analysis of the same survey experiment in Friedman et al. (2017) suggests this is due to happenstance of off-setting errors; negative errors in incidence multiplied by positive errors in value.

down to 23,000, or just six transactions per household per day (Figure 1). There was no 'bundling' into fewer, larger, transactions; for example, by reporting a composite category like "groceries" rather than individual items, and in fact the value of the average transaction fell slightly, from K4.60 to K4.00. Consequently, the 3.4 percent average daily decline in the transaction count converts into a decline in the value of daily transactions of 4.4% per day. With consumption apparently declining over time, households look poorer, the longer they are observed; the headcount poverty rate using just the second week of diary data is 47%, compared to 41% with just the first week of data (Gibson, 2013).

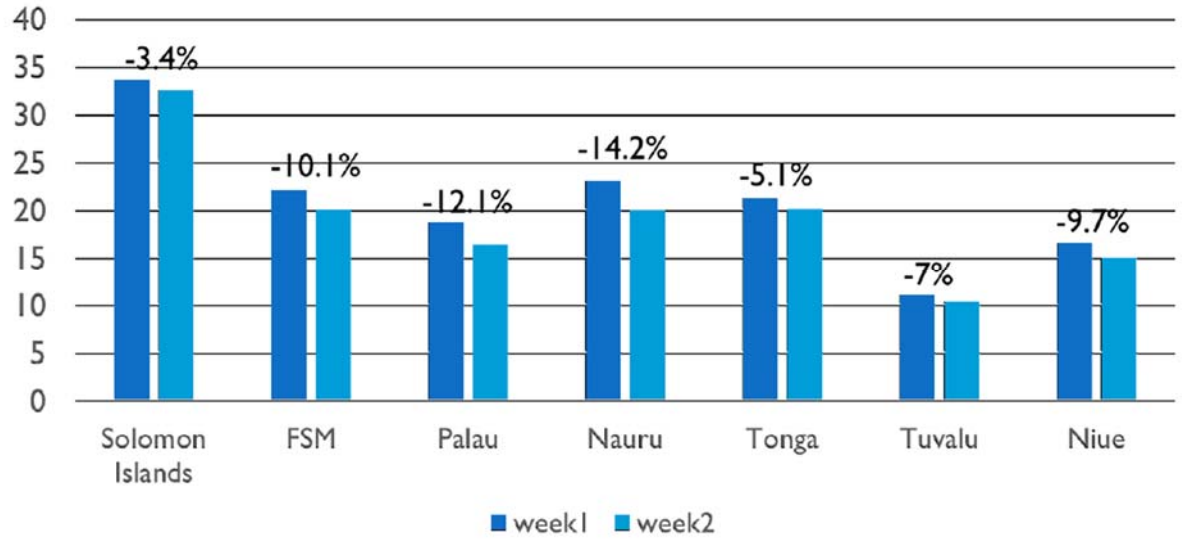
Figure 1: Diary fatigue in the 2009/10 PNG HIES



Source: Gibson 2013

Similar but not as dramatic patterns have been found across other HIES in the Pacific. Figure 2 presents the contrast between the number of food transactions in week 1 versus week 2, for a series of HIES in the Pacific. The number of food transactions recorded in diaries declines by a median of ten percent between week 1 and week 2 (Sharp, 2018).

Figure 2: Average number of food transactions reported in diary week 1 versus diary week 2



Source: Sharp et al. 2018

There is also evidence from the Pacific on impacts of attempting to measure starting and ending food stocks. Stocks are conceptually important, if indirectly deriving food consumption from an acquisitions-based survey, like a diary-keeping HIES. But measures that require looking into pantries, store rooms, and refrigerators are highly intrusive, and when overlaid with already declining compliance from the effort spent on the diaries, means that ending stocks are likely to be poorly measured, compared to starting stocks. Consequently, even though agriculture in the Pacific is largely non-seasonal, and surveys are staggered over all months of the year, so there should be no net destocking, on average, the stock measurement approach provides a – likely erroneous – net contribution to apparent food availability. For example, the 2009/10 PNG HIES measured stocks of over 100 items, and apparent destocking added 6% to the value of food consumption, and including the apparent consumption contribution from net destocking caused the headcount poverty rate to drop by four percentage points (Gibson, 2013). Likewise, in the 2012/13 HIES for the Solomon Islands, the calorie totals that included stock measurements were 6% higher, on average, with apparent net destocking adding 170 calories per person per day.

3. Experimental Design, Survey Conduct and Costs

The fieldwork for the experiment was conducted by EPPSO between July and October 2018, in four three-week rounds, with one week of rest and travel between each round. The sample design stratified households into three geographic areas (the capital, and largest urban area, Majuro; the next largest urban area, based in the Kwajalein atoll, Ebeye; and the outer islands).[4] There was one survey team per area, and teams consisted of one supervisor and five interviewers. All five treatment arms were fielded in the same area at the same time, and interviewers rotated between treatment arms across the rounds of fieldwork (so each interviewer implemented four of the five survey modules). The target sample for each arm of the experiment was based on the feasible workload for interviewers; for highly monitored diary surveys, involving seven visits to each household, an interviewer could only complete six households in the 21-day cycle, while for the less monitored diaries and for the recall modules a single interviewer could cover 18 households. The fieldwork schedules are shown in Appendix Table 1. Feedback from the interviewers was that they struggled to complete their workload for highly monitored diaries while they had ample free time during the survey rounds when they were allocated to the recall modules.

The combination of three geographic strata, four survey rounds and workloads of either six or 18 households per interviewer per round should have yielded sample sizes of either 72 (for highly monitored diaries) or 216 per arm of the experiment. In fact, the achieved sample size, and even more so for the effective sample size that only considers households with analyzable data, varied widely. Table 1 summarizes the different stages of moving from target sample size to the number of households with analyzable records. The effective completion rate is just over

[4] In Majuro and Kwajalein, EAs were selected randomly with probability proportional to size (amongst EAs with more than 80 households counted during the 2011 census). In outer atolls the most populated EAs of Ailinglaplap, Namdrick, Jaluit, and Wotje were selected. A listing operation was carried out in each selected EA prior to survey fieldwork, and households were randomly selected and randomly allocated to interviewers and survey arms.

two-thirds (68.5%) for less monitored diaries, while for the other four modules it ranges between 80-88%. These completion rates include replacements; if a selected household refused to be interviewed a replacement household was drawn randomly from a list of reserves, with the replacement rates ranging from 17.3% to 19.9% across arms. There are a few households which, due to interviewer or supervisor error, were selected but not from the replacement list (which is why Arm 1 exceeded the target sample). Even with these replacements, the completed sample was only 90% of the target sample due to refusals later in the survey process.

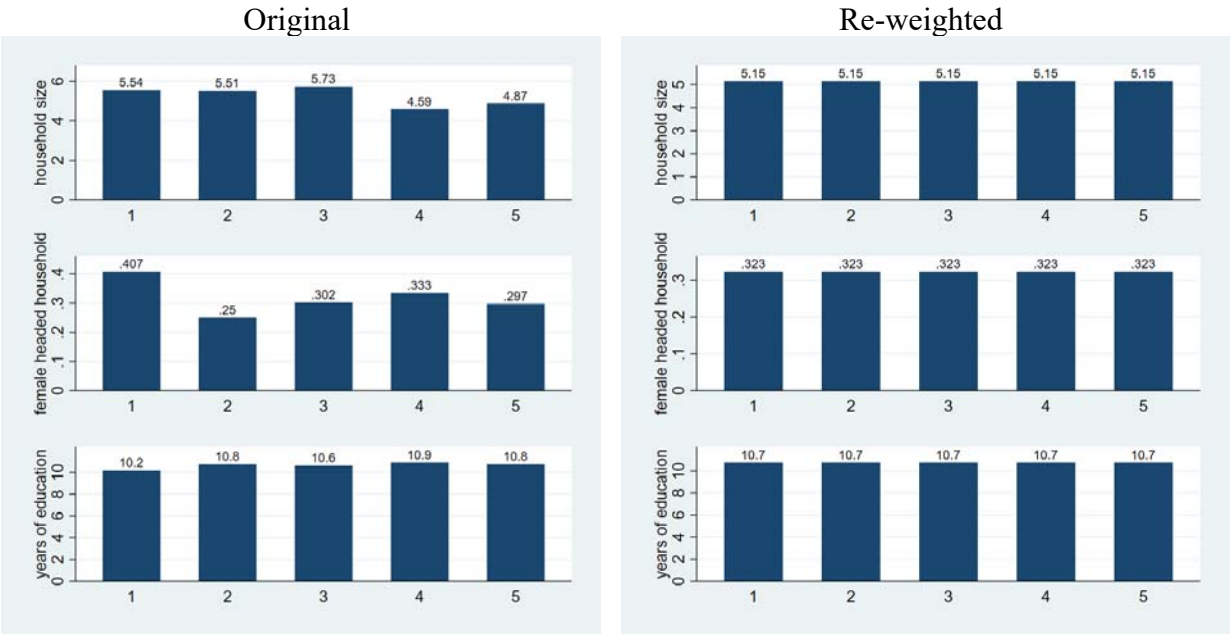
Furthermore, even for interviewed households (counted in the completed sample size of n=716), over 11% of them reported no food consumption. This was especially common with less monitored diaries, where 18% of respondents record no transactions for food consumption.[5] However, even the recall surveys had 7-10% rates of data being recorded for some sections of the questionnaire but not providing any food consumption data. In principle, a CAPI system can force interviewers to probe when respondents do not report any food consumption but here the interviewers had similar latitude as occurs under field conditions for surveys implemented by statistics offices. Thus, in contrast to experiments by researchers that may be tightly controlled, the statistics agencies that actually carry out HIES may not have the same monitoring capacity and so this experiment provides us with 'warts and all' data. These incomplete records matter to both sample balance (which is lost, even though the original selections were randomized) and to surveys costs as households without consumption data cannot be used for poverty analysis.

In terms of sample balance, for the n=634 households with analysable data, we compared various demographic characteristics between the experiment arms. Five examples are given in Appendix 2, for household size, percentage of households with female heads, years of education

[5] This is a common feature of diary-keeping surveys. For example, the first Integrated Household Survey in Malawi was administered to 13,000 households, and only 6,600 had complete expenditure information in the diaries.

of the head, the age composition of households and the employment type for the head. The average size tends to be smaller in the recall arms, fewer female-headed households respond to the less monitored diary, and the recall surveys tend to have higher numbers of working age adults (age 15 – 59) compared to those under 15 or 60 and older. Due to the refusals and incomplete surveys, the analyzable samples are not evenly distributed across the arms by geography, but using multivariate analysis shows the differences in demographic characteristics by arm hold even when controlling for location. One implication of this is that weighting for non-response (and also for incomplete data) by geographic strata does yield balance across the demographic characteristics (and so threatens the validity of estimated differences in mean consumption by arm). Instead, we use raking weights to force the following demographic characteristics to align (household size, head literacy, gender of head, years of education for head, employment categories for the head, and the consumption components that are gathered using methods that do not vary by arm, such as imputed rent and the use value of durables).

Figure 3: Illustrating the Impact of the Raking Weights for Selected Demographic Variables



Turning from sampling issues to the actual content of the modules, for the three diary arms in this experiment, acquisitions of food and frequently purchased items, such as personal care products, were collected via household-level expenditure diaries over 14 days (that is, there was one diary per household rather than separate diaries for each adult). The interviewers also measured opening and closing stocks for 28 food groups (the variety of fish and seafood stocks was so great that this group was further split into 30 categories). Other items, including expenses related to utilities, transportation, health, education, communication, and recreation, are collected via recall, using reference periods of either three months or 12 months, depending on the type of expense. For the recall arms, all expenses are collected via recall, with a 7-day recall period for food and other frequent items, and the same three month or 12 month recall period that was used for infrequent items with the diary arms. Imputed rent and the use value of durables are calculated from information collected in an identical fashion across all five arms (and thus should be invariant to the type of module used).

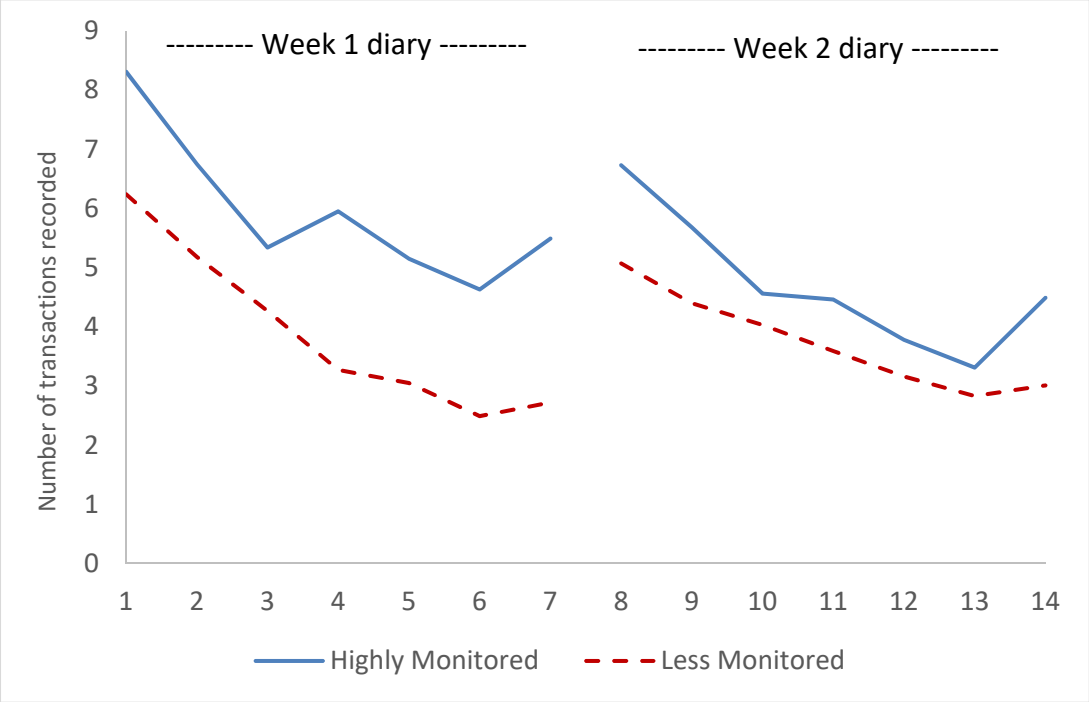
We calculated the cost of fielding each type of module in two different ways. The first was based on the actual survey budget and shows what the various costs would be if the survey had exclusively used one module type (rather than a mix of cheaper and more costly modules in operation at the same time). In other words, it is a guide to what a country with a similar cost structure to RMI would need to budget for a survey using a particular type of module. The costs are calculated separately for rural and urban sectors because rural EAs had higher transport and labour costs, partly due to the need to send urban interviewers out to rural areas given that suitable personnel are hard to recruit from rural areas (Table 2). Survey costs in RMI are high, partly due to the need for air fares and boat travel for outer atolls, and so to provide a more transferrable estimate of relative costs that may apply in other contexts, we also calculated a

more stylized budget based on variable labour costs, per-visit travel costs, printing, coding and data entry (Appendix Table 3). The two approaches yielded relative cost ratios that are similar, with highly monitored diaries costing almost five times as much as a single-visit recall (with ratios of 4.2 in urban areas and 5.2 in rural areas, as rural revisits are more costly in terms of time and transport). If highly monitored diaries are combined with CAPI, costs are even higher, as tablets cost more than paper forms and also due to a higher time demand on interviewers. The less monitored diaries are at least twice as expensive as a single visit recall, while a two-visit recall survey would be about 30% more costly than a single-visit survey. These cost estimates are all based on the completed households (n=716) rather than the households with analysable data (n=634) because it is not until all the costs have been borne that we typically know whether the records for a household are usable or not (and they may be usable for some purposes that do not involve consumption measurement and poverty analysis).

4. Results

The two PAPI diaries, some combination of which represents the *status quo* in the Pacific, show considerable diary fatigue. In the highly monitored diaries, the number of transactions listed in the diary declines from 8.3 on day 1 to 4.5 on day 14, with small jumps on the last day of each diary-keeping week (the first day is staggered over days of the week and we show in Appendix 4 that a similar decline in the number of transactions occurs, controlling for day of the week and for location). With the less monitored diaries, the decline in the number of transactions is from 6.2 on day 1 to 3.0 on day 14 (Figure 4). Combining the data from the highly monitored and less monitored diaries, the trend rate of decline is 3.4% fewer transactions recorded per day, the same rate of diary fatigue seen in the 2009/10 PNG HIES (as shown in Figure 1). Controlling for this trend, there are 35% fewer transactions recorded, on average, in the less monitored diaries.

Figure 4: Average Number of Transactions per Household per Day With Pen and Paper Diaries (by Monitoring Intensity)



While the diary fatigue reduces measured consumption, an off-setting error that raises apparent food consumption comes from stock measurement. Of 270 diary-keeping households with analyzable results, 236 reported starting food stocks but only 211 reported ending food stocks. Moreover, of those reporting food stocks at both the start and end, twice as many report larger starting food stocks than ending food stocks. The combination of these two patterns sees apparent destocking of food being equivalent to about four percent of total expenditure. For the less monitored diary, apparent destocking contributes almost 400 calories per person per day, while it contributes about 200 calories per person per day for the highly monitored diary (and about 130 calories per person per day for the CAPI diary). As noted above, there is no reason to expect net destocking in these non-seasonal environments and the most plausible explanation for these patterns is that cooperation with the measurement of food stocks is much less at the end of the 14 days of diary-keeping than it was at the start, making it appear that there has been a net destocking. This issue does not affect recall modules, which directly ask about food consumed,

rather than needing to indirectly derive food consumption from a complete accounting of inflows and outflows into the household (where net destocking counts as an inflow).

We calculated a per capita consumption estimate that was made up of four components: food and frequent non-food expenses that were either recorded in diaries or reported by recall (annualizing from the 14-day or 7-day periods); infrequent expenses that were only obtained by recall, over either 3-months or 12-months; items such as alcohol and tobacco whose value of consumption was obtained from both diary and recall; and, imputed rents and the use value of durables that are based on calculations (and should not vary by arm as the variables supporting the calculations are asked in identical ways for all households).

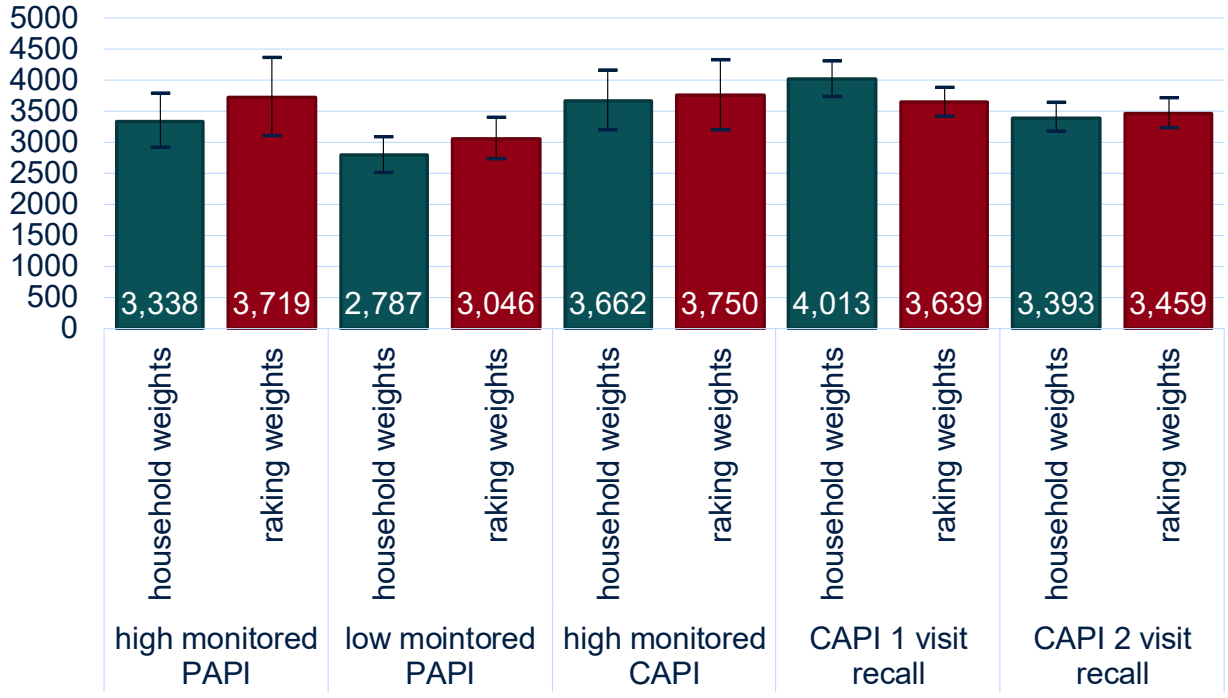
In Figure 5 we show these components, and estimated total per capita consumption, by experiment arm (where the averages rely on the raking weights that restore arm-by-arm balance over demographic characteristics). For food and frequent non-food, which is the most impacted by the variations in survey module design, the less monitored diary yields average consumption values that are just 61% of what the highly monitored diaries yield (and 67% of what single visit recall yields). The interpretation of this likely understatement of food and frequent non-food consumption, when using less monitored diaries, is colored by this module having the largest upward bias from apparent destocking of food; thus, in surveys where food stocks are not being measured, the likely understatement of actual consumption when using less monitored diaries would be even greater. For the two highly monitored diaries, the average value of food and frequent non-food consumption ($1247) is identical to what single-visit recall yields ($1248) and is just slightly higher than the average across the two types of recall ($1210). The infrequent expenses, which are gathered with the same recall method over all arms, show much less variation, averaging $1220 across the two types of recall and $1120 across the three types of

diaries, while the calculated component of consumption (for imputed rent and the use value of durables) does not vary by arm. Putting all the components together, per capita total consumption would average $3046 per person per year when using less monitored diaries, just 82% of what it averages with highly monitored diaries and 84% of what single-visit recall yields. For the two highly monitored diaries, average total per capita consumption would be $3735, which is just 2.6% above the average that one gets from using a much cheaper single visit recall survey.

Figure 5: Per Capita Consumption Components and Totals by Experiment Arm

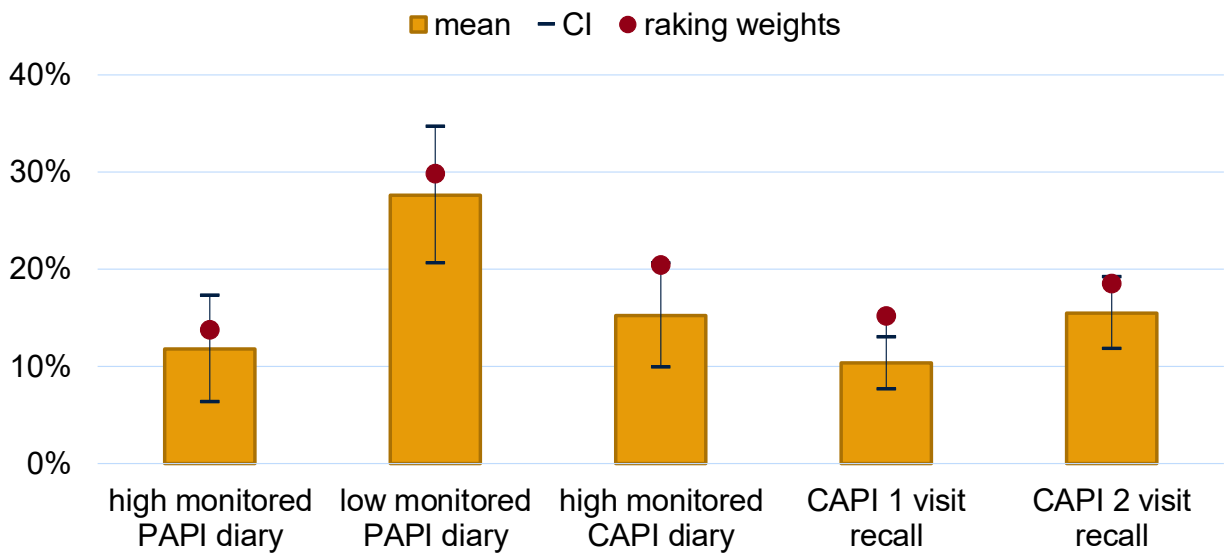
We show the estimates of average per capita total consumption, and confidence intervals, in Figure 6. The averages are presented using two types of weights: the household weights that do not restore the balance lost by differential refusals and incomplete data by module type, and the raking weights designed to restore balance. If we only used the household weights, the differences between arms would seem larger; the total per capita consumption with the less monitored diary would average only 69% of what single visit recall yields (and 83% of what the more monitored PAPI diary yields) and the average of the two more monitored diaries would be five percent below the average of the two recall modules. However, the general pattern of the results would be the same, which is that less monitored diaries yield lower estimates of total consumption than from any other module, and that, notwithstanding the much greater cost of data collection, the highly monitored diaries yield average consumption values that are similar to what we get using recall.

Figure 6: Average Per Capita Total Consumption by Experiment Arm: Using Two Types of Weights



Finally, we consider the implications of these measurement differences for poverty analysis. In Figure 7 we show the estimated incidence of poverty under a relative poverty line set as half the median real per capita consumption. We show the poverty rates using both the raking weights and the household weights, and they tell a similar story, which is that about 30% of the population would appear to be living in households below the poverty line if a less monitored diary is used to gather consumption data. In contrast, the poverty rates are from 10-15% (or 15-20% if using raking weights) for all of the other modules. The second notable result in Figure 7 is that, despite the much greater cost of data collection, the poverty rate with highly monitored diaries (either PAPI or CAPI) is not much different to what would be found with a much cheaper recall survey.

Figure 7: Poverty Rates by Experiment Arm: Relative Poverty Line as Half the Median



5. Conclusions

In the Pacific, as in much of the rest of the world, Household Income and Expenditure Surveys have to provide date for multiple uses. Historically, some of those uses, such as updating CPI weights, have swung the balance towards using diary-keeping expenditure surveys, because of the finer detail on commodities and outlets that may be obtained with diaries (even though some

CPIs have managed with just recall surveys). Unfortunately, diary-keeping surveys are very expensive to field, and very burdensome for respondents, and they are poorly suited as a data source of poverty monitoring.

The first problem is that the high cost of diary-keeping surveys limits their frequency – to just once every nine years in the Pacific, on average – and the time-consuming fieldwork, coding and data entry limits their timeliness. The second problem is that the burden of these surveys leads to high rates of non-response, which reduces power, raises cost per completed interview, and may impair the representativeness of their results. A third, and related, problem is that the survey, as implemented, will vary even at the household level, in terms of monitoring – thus the end result will be somewhere between the highly monitored and the less monitored diary that we have experimented with here. Given that the overall performance of diary-keeping surveys will be some mix of what we have studied here, they can be expected to have low effective completion rates, and even for records that provide some consumption data there may be considerable understatement as respondents progressively lose enthusiasm and only partially record transactions. The overall rate of decline in reported transactions with diaries that we find, of 3.4% per day, means that households look poorer the longer we observe them and this complicates poverty analysis.

Our first key finding is that the overall performance of diary surveys is worse than the performance of recall surveys, given that there is a chance that the diary-keeping survey will devolve into the less monitored variant. Our second key finding is that even if statistics offices could guarantee that they always use highly monitored diaries, we can get similar information on average consumption and on poverty rates, using far cheaper recall surveys. These findings provide a basis for hope that regions like the Pacific, where some countries are far away from

meeting 2030 poverty goals, could develop a much more effective poverty monitoring survey infrastructure that provided more frequent and more timely data if they would move away from their tradition of relying on diary-keeping surveys.

References

Backiny-Yetna P, Steele D, Yacoubou Djima I. 2014. The Impact of Household Food Consumption Data Collection Methods on Poverty and Inequality Measures in Niger. Policy Research Working Paper; No. 7090. World Bank Group, Washington, DC. World Bank. https://openknowledge.worldbank.org/handle/10986/20626

Bee, A., Meyer, B.D. and Sullivan, J.X., 2012. *The validity of consumption data: are the consumer expenditure interview and diary surveys informative?* (No. w18308). National Bureau of Economic Research.

Beegle K, De Weerdt J, Friedman J, Gibson J. 2012. Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania. *Journal of Development Economics* 98(1): 19–33.

Brzozowski, M., Crossley, T.F. and Winter, J.K., 2017. A comparison of recall and diary food expenditure data. *Food policy*, *72*, pp.53-61.

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

de Nicola F, Giné X, 2014. How accurate are recall data? Evidence from coastal India. *Journal of Development Economics*, *106*, pp.52-65.

De Weerdt, J, Beegle, K, Friedman, J, Gibson, J. 2016. The challenge of measuring hunger through survey. *Economic Development and Cultural Change* 64(4): 727-758.

Dupriez, O, Smith, L, Troubat, N. 2014. *Assessment of the Reliability and Relevance of the Food Data Collected in National Household Consumption and Expenditure Surveys*. FAO, IHSN and World Bank accessed 6 April 2014: http://www.ihsn.org/home/node/34

Eisenhower, D., Mathiowetz, N.A. and Morganstein, D., 2004. Recall error: Sources and bias reduction techniques. *Measurement errors in surveys*, pp.125-144.

Fiedler J. 2013. Towards Overcoming the Food Consumption Information Gap: Strengthening Household Consumption and Information Surveys for Food and Nutrition Policymaking. *Global Food Security* 2(1): 56-63.

Fiedler, J, Yadav, S. 2017. How can we better capture food away from home? Lessons from India's linking person-level meal and household-level food data. *Food Policy* 72(1): 81-93.

Friedman, J, Beegle, K, De Weerdt, J, Gibson, J. 2017. Decomposing response error in food consumption measurement: Implications for survey design from a randomized survey experiment in Tanzania. *Food Policy* 72(1): 94-111.

Gaskell G, Wright D, O'Muircheartaigh C. 2000. Telescoping of Landmark Events: Implications for Survey Research. *The Public Opinion Quarterly, 64*(1), 77-89. Retrieved from http://www.jstor.org.libproxy-wb.imf.org/stable/3078842

Gibson, J. 2013. Two decades of poverty in Papua New Guinea. Presentation at the Crawford School, Australian National University, Canberra.

Gibson, J., Kim, B. 2012. Testing the infrequent purchases model using direct measurement of hidden consumption from food stocks. *American Journal of Agricultural Economics* 94(1): 257-270.

Gieseman, R., 1987. The Consumer Expenditure Survey: quality control by comparative analysis. *Monthly Lab. Rev.*, *110*, p.8.

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.

Loftus E, Marburger W. 1983. Since the Eruption of Mt. St. Helens, Has Anyone Beaten You Up? Improving Accuracy of Retrospective Reports with Landmark Events. *Memory and Cognition* 11(2): 114-20.

Morwitz, V.G., 1997. It seems like only yesterday: The nature and consequences of telescoping errors in marketing research. *Journal of Consumer Psychology*, *6*(1), pp.1-29.

Neter, J. and Waksberg, J., 1964. A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, *59*(305), pp.18-55.

Schündeln, M. 2018. Multiple Visits and Data Quality in Household Surveys. *Oxford Bulletin of Economics and Statistics*. 80(2): 380-405.

Scott, C., Amenuvegbe, B. 1991. Recall loss and recall duration: an experimental study in Ghana. *Inter-Stat,* 4(1): 31-55.

Sharp M, Buffiere B, Menaouer O. 2018. Household Income and Expenditure Surveys (HIES) in the Pacific Region, paper presented to Pacific Statistics Methods Board, Nadi, Fiji, October 2018. Viewed 9 April, 2019. https://sdd.spc.int/images/documents/Meetings/Methods_Board/30-31_Oct_2018/PSMB2_2018_Doc1_Sharp_et_al_HIES.pdf

Zezza, A., Carletto, C., Fiedler, J.L., Gennari, P. and Jolliffe, D., 2017. Food counts. Measuring food consumption and expenditures in household consumption and expenditure surveys (HCES). Introduction to the special issue. *Food Policy*, *72*, pp.1-6.

Appendix Table 1: Fieldwork Schedule for Interviewers for the Different Arms of the Experiment

(A) Highly Monitored Diary (either PAPI or CAPI) With 7 Visits to Each Household

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20	Day 21
Visit #	0	1		2		3		3	4		5		6		6	7					
Field work	HH listing	HH 1	HH 4	HH 1	HH 4	HH 1	Rest	HH 4	HH 1	HH 4	HH 1	HH 4	HH 1	Rest	HH 4	HH 1	HH 4	Spare days to complete interviews / diary, if needed			
		HH 2	HH 5	HH 2	HH 5	HH 2		HH 5	HH 2	HH 5	HH 2	HH 5	HH 2		HH 5	HH 2	HH 5				
		HH 3	HH 6	HH 3	HH 6	HH 3		HH 6	HH 3	HH 6	HH 3	HH 6	HH 3		HH 6	HH 3	HH 6				
Daily activities																					
First contact	X																				
HIES module		X	X	X	X																
Day 0 food stock		X	X																		
Drop week 1 diary		X	X																		
Check week 1 diary				X	X	X		X	X	X											
Pick up week 1 diary									X	X											
Drop week 2 diary									X	X											
Check week 2 diary											X	X	X		X	X	X				
Pick up week 2 diary																X	X				
Day 15 food stock																X	X				
Diary data entered by EPPSO																					EPPSO
Daily data backup		X	X	X	X	X			X	X	X	X	X		X	X	X				

(B) Less Monitored Diary (PAPI only) With 3 Visits to Each Household

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20	Day 21
Visit #	0	1						2							3						
Field work	HH listing	HH 1	HH 5	HH 9	HH 13	HH 17	Rest	HH 1	HH 5	HH 9	HH 13	HH 17		Rest	HH 1	HH 5	HH 9	HH 13	HH 17		
		HH 2	HH 6	HH 10	HH 14	HH 18		HH 2	HH 6	HH 10	HH 14	HH 18			HH 2	HH 6	HH 10	HH 14	HH 18		
		HH 3	HH 7	HH 11	HH 15			HH 3	HH 7	HH 11	HH 15				HH 3	HH 7	HH 11	HH 15			
		HH 4	HH 8	HH 12	HH 16			HH 4	HH 8	HH 12	HH 16				HH 4	HH 8	HH 12	HH 16			
Daily activities																					
First contact	X																				
HIES module		X	X	X	X	X															
Day 0 food stock		X	X	X	X	X															
Drop week 1 diary		X	X	X	X	X															
Pick up week 1 diary								X	X	X	X	X									
Drop week 2 diary								X	X	X	X	X									
Pick up week 2 diary															X	X	X	X	X		
Day 15 food stock															X	X	X	X	X		
Diary data entered by EPPSO																					EPPSO
Daily data backup		X	X	X	X	X		X	X	X	X	X			X	X	X	X	X		

(C) Single Visit Recall (CAPI only)

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20	Day 21
Visit #		1	1	1	1	1		1	1	1	1										
Field work	HH listing	HH 1	HH 3	HH 5	HH 7	HH 9	Rest	HH 11	HH 13	HH 15	HH 17	If required, finish off interviews for allocated work load									
		HH 2	HH 4	HH 6	HH 8	HH 10		HH 12	HH 14	HH 16	HH 18	Help other team members to complete their allocated work load									
Daily activities																					
First contact	X																				
Complete interviews		X	X	X	X	X		X	X	X	X										
Daily data backup		X	X	X	X	X		X	X	X	X										

(D) Two Visit Recall (CAPI only)

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15	Day 16	Day 17	Day 18	Day 19	Day 20	Day 21
Visit #		1	1	1	1	1		1	1	1	1										
Field work		HH 1	HH 3	HH 5	HH 7	HH 9	Rest	HH 11	HH 13	HH 15	HH 17										
		HH 2	HH 4	HH 6	HH 8	HH 10		HH 12	HH 14	HH 16	HH 18										
Visit #	HH listing								2	2	2	2	2	2	2	2	2	2			
Field work									HH 1	HH 3	HH 5	HH 7	HH 9	Rest	HH 11	HH 13	HH 15	HH 17			
									HH 2	HH 4	HH 6	HH 8	HH 10		HH 12	HH 14	HH 16	HH 18			
Daily activities																					
First contact	X																				
Complete HIES module		X	X	X	X	X		X	X	X	X										
Advise household that you will return		X	X	X	X	X		X	X	X	X										
Complete recall									X	X	X	X	X		X	X	X	X			
Daily data backup		X	X	X	X	X		X	X	X	X	X	X		X	X	X	X			

Appendix 2: Comparisons of Demographic Characteristics By Experiment Arm

Figure 1: Basic demographic characteristics by arm

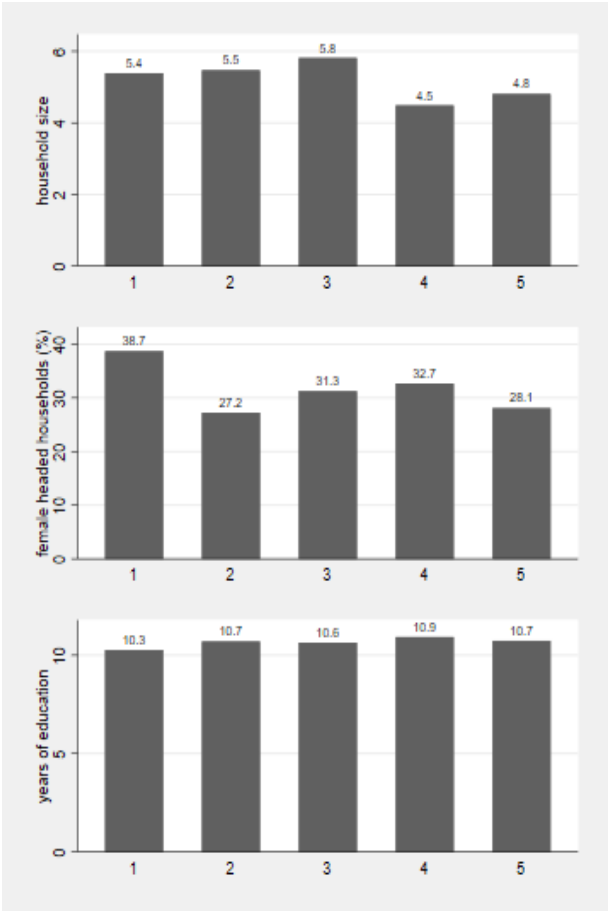


Figure 2: Household composition by arm

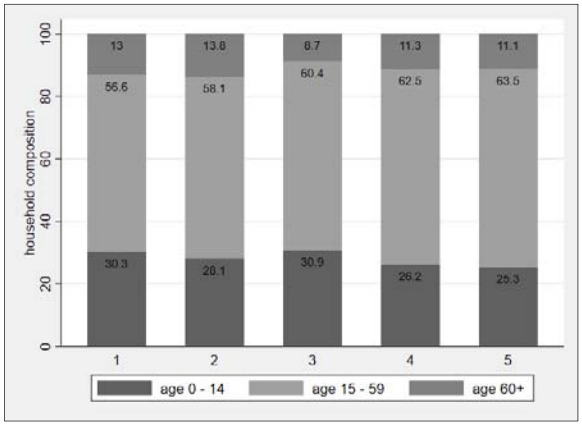
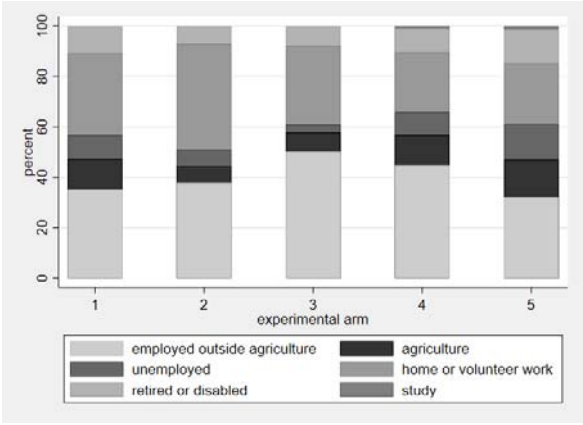


Figure 3: Employment type (head) by arm



Appendix Table 3: Stylized Budget for Variable Costs for Each Survey Module

Arm	Survey Type	HH per Round Target	Actual	nVisits	Hours/HH	Labor	Travel	Printing	Coding	Entry	Total Cost	**Ratio to Arm 4**
1	7-visit 14-day diary, PAPI	18	19	133	5	50	35	5	10	10	110	**4.4**
2	3-visit 14-day diary, PAPI	54	45	135	3	30	15	5	10	10	70	**2.8**
3	7-visit 14-day diary, CAPI	18	16	112	8	80	35	0	10	0	125	**5.0**
4	1-visit 7-day recall, CAPI	54	50	50	2	20	5	0	0	0	25	
5	2-visit 7-day recall, CAPI	54	49	98	2.5	25	10	0	0	0	35	**1.4**
		198	179	528								
	13500											
	8820											
	22320	1.65333										

Salary and *per diem* per 3-week round total $22,320 (for three teams, each with 1 supervisor and 5 interviewers and per diem paid only for the remote strata)

Interviewer labor costs $40 per day; supervisor covers 5 interviewers and costs $50 per day (so total pro-rated $50/interviewer day or $10/hour)

Hours required based on 1.5 hrs for the non-consumption, 0.5 hrs for recall, 20min for stocks (x2), 10min per diary check and 30m entry per check (arm 2 only) and 30 min scheduling

Transport cost per round is $2650, so pro-rated as $5 per visit

Printing for PAPI includes freight to RMI, total cost of NZ$1840 or $1200 (so $300 per round) or $5 per household

Salary for COICOP coding and data entry is $40 per day, and productivity for either task is 4 diaries per staff day.

Appendix 4: Linear prediction for number of diary items (controlling for atoll and day of the week)

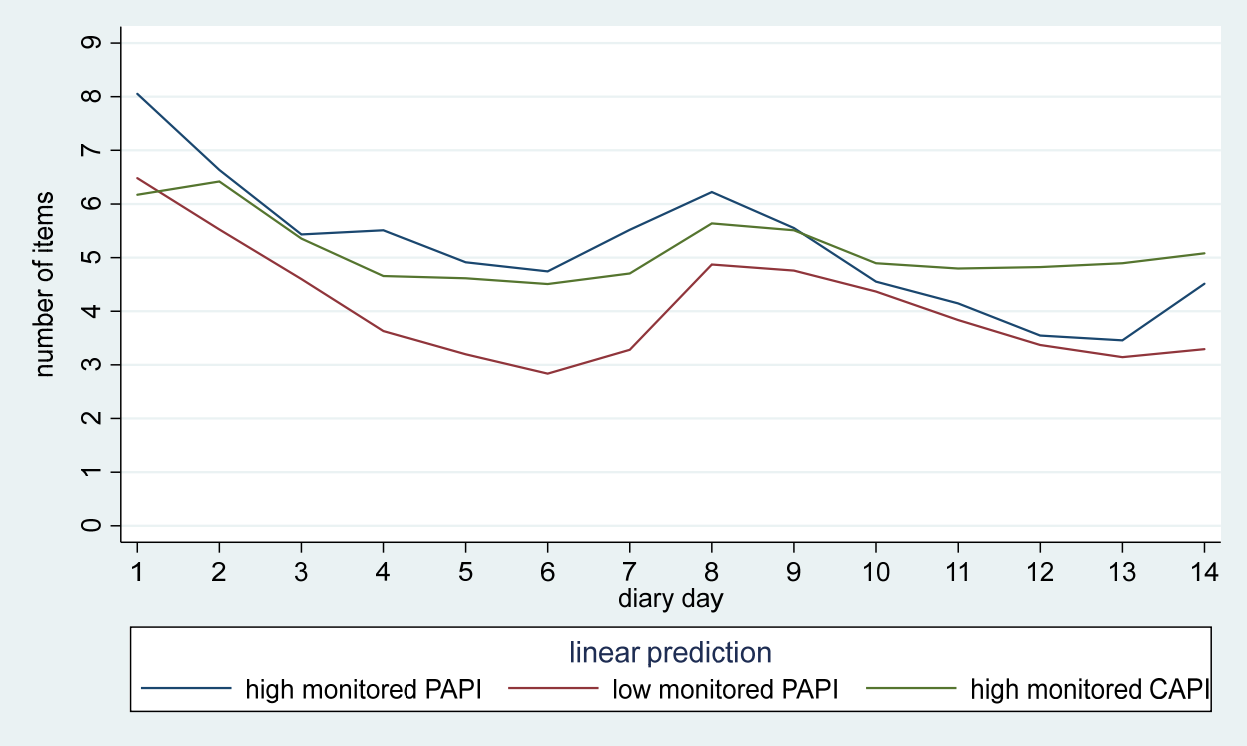


Table 1: Description of Treatment Arms and Sample Sizes

Arm	Description	Workload per interviewer	Target sample size	Replace-ment rate	Completed sample size	# with no reported food cons	# with analysable records	Effective completion rate
1	14-day diary, highly monitored (visits every 2 days), transactions recorded using pen and paper	6	72	17.3%	75	16	59	81.9%
2	14-day diary, less monitored (visits each week), transactions recorded using pen and paper	18	216	19.9%	181	33	148	68.5%
3	14-day diary, highly monitored (visits every 2 days), data entered by interviewer using CAPI during each visit	6	72	18.8%	64	1	63	87.5%
4	7-day single visit recall, for list of 102 food groups and 20 non-food groups, CAPI data entry during the interview	18	216	18.6%	199	19	180	83.3%
5	7-day two-visit recall, using list of 102 food groups and 20 non-food groups, CAPI data entry during the interviews	18	216	19.3%	197	13	184	85.2%
Total			792		716	82	634	80.1%

Table 2: Total Costs per Completed Interview

			number of interviews												
Arm	Survey Type	Location	Target	completion rate	Actual	Number of days in field	Fixed costs for survey	Fixed costs for arm	Labor per arm	Fixed Transport per arm	Variable transport per arm	coding & entry	Total cost per cluster	Cost per completed interview	Indexed
1	7-visit 14-day diary, PAPI	urban	6	1.0416667	6.25	18	2288.18	149.50	3981.26		1080	125	7624	1220	4.22
		rural	6	1.0416667	6.25	21	2288.18	149.50	4644.81	2400	9240	125	18847	3016	10.43
2	3-visit 14-day diary, PAPI	urban	18	0.7986111	14.375	18	2288.18	280.83	3981.26		1080	287.5	7918	551	1.91
		rural	18	0.9166667	16.5	21	2288.18	280.83	4644.81	2400	9240	330	19184	1163	4.02
3	7-visit 14-day diary, CAPI	urban	6	0.8333333	5	18	2288.18	931.50	3981.26		1080	50	8331	1666	5.76
		rural	6	1	6	21	2288.18	931.50	4644.81	2400	9240	60	19564	3261	11.28
4	1-visit 7-day recall, CAPI	urban	18	0.8888889	16	5	2288.18	931.50	1105.91		300		4626	289	1.00
		rural	18	0.9861111	17.75	7	2288.18	931.50	1548.27	2400	3080		10248	577	2.00
5	2-visit 7-day recall, CAPI	urban	18	0.8819444	15.875	10	2288.18	931.50	2211.81		600		6031	380	1.31
		rural	18	0.9722222	17.5	14	2288.18	931.50	3096.54	2400	6160		14876	850	2.94

Per cluster (assumes teams of 5+1 supervisor)
"Fixed costs" are total fixed costs / 12 clusters
"Fixed costs per arm" include printing and transport
* sat phone excluded