Policy Research Working Paper 10029

Effects of Data Collection Methods on Estimated Household Consumption and Survey Costs

Evidence from an Experiment in the Marshall Islands

Michael K. Sharp
Bertrand Buffière
Kristen Himelein
Nathalie Troubat
John Gibson

WORLD BANK GROUP
Poverty and Equity Global Practice
April 2022

Policy Research Working Paper 10029

Abstract

In the Pacific, multitopic household surveys have historically gathered expenditure data using open form diaries completed on paper. This methodology is costly to governments, is burdensome for respondents, and takes substantial time to process the results. Noncompliance and partial compliance in diary keeping can artificially inflate poverty measures, biasing economic statistics. This paper reports findings from an experiment in the Marshall Islands comparing the cost and accuracy of several collection methodologies. Variable costs for the status quo diary survey design are between 2.8 and 4.4 times more expensive than a single-visit seven-day recall survey, with the tablet-based diary being even more costly. The highly monitored diaries give similar results to recall but at much greater cost; the status quo yields data of worse quality as effective completion rates with low monitored diaries are only two-thirds the completion rates of recall-based options. Finally, the paper discusses the implementation challenges associated with the different methods in a capacity-constrained environment.

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

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

Michael K. Sharp, Bertrand Buffière, Kristen Himelein, Nathalie Troubat, and John Gibson[1]

JEL: C81, O15, Q18

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

Acknowledgments: Economic Policy, Planning and Statistics Office of the Government of the Marshall Islands for hosting the experiment. Funding provided by the Government of the Marshall Islands, the Australia Department for Finance and Trade, the New Zealand Ministry of Foreign Affairs and Trade and the Australian Centre for International Agricultural Research (FIS 2018/155). The authors are grateful to comments from the Pacific Statistics Methods Board, participants at the 2019 IARIW-World Bank conference, Kathleen Beegle, David McKenzie, and Neil Andrew. The authors acknowledge the late Pierre Wong for his significant role in development of consumption surveys in the Pacific region. All remaining errors are those of the authors.

[1] Michael Sharp is with the Statistics for Development Division (SDD) of the Pacific Community (SPC) and the University of Wollongong and is the corresponding author (michaels@spc.int). Bertrand Buffière is with SDD-SPC, Kristen Himelein is with the Poverty and Equity Global Practice of the World Bank, Nathalie Troubat is a Food Security Specialist working with SDD-SPC, and John Gibson is with the Department of Economics, University of Waikato. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the Pacific Community, the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

1. Introduction

In much of the world, data from multitopic household surveys, like the Household Income and Expenditure Surveys (HIES), serve multiple purposes, including poverty and food security monitoring, updating the CPI basket, and as an input to National Account calculations. Given their length and complexity, these surveys are very costly to field, particularly in the Pacific Islands, which are characterized by sparse populations and high transportation costs. Recent HIES had per completed household costs ranging from almost US$700 in Solomon Islands to US$4,000 per household in Papua New Guinea (PNG). Even in less challenging environments, such as the smaller atoll countries, costs were up to US$800 per surveyed household. These high costs were driven mainly by the need for interviewers to remain in villages for 3 weeks to oversee 14-day diaries for consumption data, and meant that HIES were fielded only infrequently, with an average gap of nine years between each HIES in the region. These gaps, coupled with the time required to clean and process diary data, limited the usefulness of data from these surveys for program monitoring and policy making.

The method of collection also has implications for data quality. Diary-keeping surveys are burdensome for respondents, potentially leading to bias from diary fatigue. Noncompliance and partial compliance due to fatigue complicates analyses because it is not possible 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 produced 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 percent versus 47 percent if only the second seven days are used (Gibson, 2013). Even when no ambiguity exists, such as those cases in which no consumption is recorded for several days, the household must be dropped from the data set, a costly loss in a region with high data collection costs.

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 designed to understand the cost and data quality implications around the choice of survey mode (paper or electronic) and the methodology used for collecting consumption and expenditure data through a

household survey. 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 done by 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 in the Pacific, but the lessons learned are applicable generally to contexts with limited capacity and high per-survey costs. Additionally, while the recent survey methodology literature from both developed and developing countries has generally concluded recall data collection to be more effective than traditional diaries (Zezza et al, 2017; FAO and The World Bank, 2018), 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, fruit, and tubers.

The experiment has five arms, each fielded within the same enumeration areas at the same time, which included different combinations of paper or electronic survey mode, high and low supervision, diary and recall, representing the common choices facing NSOs in the Pacific and globally. Interviews were conducted by interviewers from EPPSO as to replicate real-world conditions and implementation challenges. Though the choice of diary versus recall has implications on sampling design, with recall surveys requiring less time in each location, and therefore being more conducive to smaller cluster sizes leading to lower design effects and higher precision for a given sample size, the focus of this paper is mainly on non-sampling error. Following the Weisberg (2005) framework on total survey error, we consider non-sampling error in respondent selection and accuracy issues, as well as the vulnerability to administration issues of five treatment arms.

The remainder of the paper is as follows: Section 2 briefly reviews related literature and previous evidence from the Pacific, Section 3 describes the design of the experiment, and Section 4 presents results on ease of implementation, data quality, and cost. Section 5 concludes with discussion, recommendations for future data collection, and areas for further research.

2. Literature review and local context

2.1. Literature review

While recall data asks respondents to remember, diaries are in principle collected in real time, and therefore, absent recall error, a well-implemented diary should provide more accurate results and better granularity than recall data. The field of survey methodology has generated many studies that examine the conditions under which recall error impacts data quality; see Sudman and Bradburn (1974) and Dex (1995) for summaries of this literature. 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 (backward 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 (two-visit) and unbounded (one-visit) recall periods because bounding has been found to decrease telescoping and increase accuracy (Loftus and Marburger, 1983); however, implementation error for the two-visit recall precluded a proper test of bounding effects.

The third source of error, omission, refers to excluding relevant events that 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 average daily expenditures reported by survey respondents in Ghana fell by almost 3 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 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 in Tanzania 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 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 shortcomings in both diary and recall measures. 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. 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." Guidelines prepared by the Inter-Agency and Expert Group on Food Security, Agricultural and Rural Statistics (FAO and World Bank, 2018) for collecting food data in consumption and expenditure surveys in low- and middle-income countries also recommended food consumption data be collected via 7-day recall but called for further research in methods to estimate consumption of food away from home (FAFH), which refers to consumption of food that is both acquired and consumed away from the dwelling, such as in a restaurant, at a school feeding program, or non-alcoholic beverages and snacks consumed at the workplace.

2.2. *Prior evidence from the region*

Historically, HIES carried out in the Pacific region prior to the current experiment gathered expenditure data using open form diaries completed on paper. As such, evidence from the region mainly deals with the shortcomings of diary collections. 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 to check on the diary-keeping households approximately every second day. The diary-keeping was staggered, both over the months of the year and the days of the week. Across the 3,800 households who provided diary data, from a target of 4,100, a total of 37,000 transactions were recorded on the first day of diary-keeping or 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 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 percent 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 percent, compared to 41 percent with just the first week of data (Gibson, 2013). Similar though less dramatic patterns have been found across other HIES in the Pacific. Sharp (2018) found

the number of food transactions recorded in diaries declines by a median of 10 percent between week 1 and week 2 for a series of HIES in the Pacific (Figure 2).

Food consumption data have also historically been collected in the Pacific without a specific module dedicated to consumption of FAFH or prepared food away from the dwelling. While reliable statistics on the share of consumption spent on FAFH are not yet available, the survey literature has shown substantial shares were common in both developed and developing country contexts and that these amounts were increasing with income (Claro et al, 2014, Smith et al, 2014, Farfán et al, 2017, De Brauw and Herskowitz, 2021, among others). Despite the importance of FAFH to an accurate understanding of health and financial well-being, a global review of questionnaires by Smith et al (2014) to assess the adequacy of FAFH measures found only 42 percent of surveys met the reliability criteria defined by the authors and nearly 25 percent of countries used only one question to capture all meals for all members.

Another important feature of diaries is that they capture the amount of food and non-food items acquired during the collection period but cannot directly measure consumption. Instead, diaries indirectly derive consumption as a residual from the following components:

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

As a result of this indirect approach, interviewers must attempt to measure opening and closing food stocks because solely using acquisitions-based diaries may either understate the food available, and by extension the value of food consumption, if the household consumes from existing stocks or else overstate if acquisitions during the diary-keeping period go into ending stocks. The impact of stocks is particularly relevant for the Pacific, which features non-seasonal agriculture, bulky root crops, high transactions costs of going to markets and gardens, and irregular shipments to remote locations, which results in considerable in-home storage and hidden consumption from stocks (Gibson and Kim, 2012). However, measures that require looking into pantries, storerooms, and refrigerators are

highly intrusive, and when overlaid with already declining compliance from the effort spent on the diaries, this 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 percent 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 percent higher, on average, with apparent net destocking adding 170 calories per person per day.

3. Experimental design

3.1. Survey design

The content of the surveys used in the experiment mimicked those of a standard HIES survey and included questions on household composition, member demographic and employment information, dwelling characteristics, etc., in addition to consumption and expenditure information. The consumption and expenditure information was collected differently across the treatment arms and included variation on recall method (diary vs recall), supervision level (high vs low), and mode of implementation (electronic vs paper). The five survey design arms were:

1. Arm 1: 14-day household-level 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. Arm 2: 14-day household-level 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. Arm 3: 14-day household-level diary, highly monitored (with interviewer visits every two days), with transactions transcribed onto tablets using Computer Assisted Personal Interviewing (CAPI) software during each interviewer visit,
4. Arm 4: 7-day single-visit recall, using a list of 102 food products and 20 non-food groups, with data entered on tablets using CAPI during the interview,
5. Arm 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 products and 20 non-food groups, with data entered on tablets using CAPI during the interview.

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 labor and travel for interviewer revisits, the survey can degrade into low monitored diaries, which may further devolve into a pseudo-recall survey. The experiment arms 1, 2 and 3 also included a section that collected food items in stocks and individual FAFH consumption on a daily basis over the 14-day diary period which was not included in the standard historic HIES methodology.

For the recall arms, Arms 4 and 5, expenses were collected via recall, with a 7-day recall period for food, an individually administered 7-day recall on FAFH consumption, and 7-day recall for other consumable items that are typically collected in diaries. Depending on the type of expense, the same 7-day, 1-month 3-month and 12-month recall period as the diary arms was used for non-food consumption, such as utilities, transportation, health, education, communication, and recreation. Imputed rent and the use value of durables were calculated from information collected in an identical fashion across all five arms. All common individual, dwelling characteristics and non-food consumption were collected using the same questionnaire, which was administered via face-to-face interview with the responses entered directly into a tablet-based Survey Solutions (World Bank, 2018) data entry platform.

3.2. Sample design and weighting

The sample design for the experiment included stratification by geography and by treatment arm. The geographic stratification divided the country into three areas (Majuro – the capital, and largest urban area; Ebeye – the next largest urban area, based in the Kwajalein atoll; and the rural outer islands), targeting the most populous areas of the country rather than generating nationally representative estimates. In Majuro and Kwajalein, the primary sampling units were census enumeration areas (EA) selected randomly with probability proportional to size from the universe of EAs with more than 80 households at the time of 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. The final design selected four urban EAs in both Majuro and Ebeye, as well as four EAs on the outer islands. Within EAs, households were randomly selected.

The sample size varied based on the treatment strata. Due to the time requirements of the two highly monitored diary arms (arm 1 – PAPI and arm 3 – CAPI), the targeted sample size was only 6 households per cluster per round. For the other three arms, which are much less demanding of interviewer and respondent time, the target was 18 households per cluster per round. The combination of three geographic strata, four survey rounds, and workloads of 6 or 18 households per interviewer per round should have yielded sample sizes of 72 for the two highly monitored diary arms and 216 for each of the other three arms of the experiment for a total sample size of 792.

Three types of weighting schemes were considered in the analysis. First, since probability proportional to size selection in the first stage followed by the selection of a constant number of households within an EA yields approximately self-weighting results at the stratum level, analysis was done first using no compensatory weights. Then probability weights were calculated to account for discrepancies between the frame and listing EA populations, non-response within EAs, and a calibration to known stratum-level populations totals, which is the most common approach to weighting in developing country contexts. A third set of weights was also calculated using more complex re-weighting procedures which aligned household size, literacy status of household head, sex of head, years of education for head, employment categories for the head, imputed dwelling rent, and the use value of household durable goods to the mean across the five experimental arms. The latter two categories are components of the consumption aggregate that are drawn from questions collected identically across the five treatment arms. This reweighting approach is an attempt to remove the impact of wealth effects in response rates in treatment arms rather than to generate representative estimates of the true population.

3.3. Fieldwork

The fieldwork was conducted by EPPSO between July and October 2018, in four three-week rounds, with one week of rest and travel between each round. 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 such that 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 low monitored diaries and for the recall modules a single interviewer could cover 18 households. The fieldwork schedules are shown in Table 1 in the appendix.

4. Results

4.1. Implementation challenges

During implementation, the teams encountered several difficulties in implementing the experiment as designed. Households that initially refused were replaced randomly from a list of replacements, though in some cases there were not enough replacement households available, and therefore some combinations of arm-location do not have the targeted number of observations. In addition, there were a few households which, due to interviewer or supervisor error, were selected but not from the replacement list, leading to Arm 1 exceeding the target sample (Table 2).

Beyond standard issues of participation, there were errors in how the consumption data was collected. The section below discusses completion rates at the item and survey level, but there were major issues specifically with the implementation of the two-visit recall arm. The design instructed interviewers to visit the household one week before the consumption data was to be collected to let the respondents know that they would be returning to ask about the intervening period. The actual period between visits, however, varied. Visit 2 was 8 days after visit 1 for 53 percent of households, 9 or more days after for 14 percent of households, and 7 or fewer days after for 33 percent of households. Additionally, the questionnaire for the two-visit recall arm asked about consumption in the last 7-days rather than consumption since the last visit of the enumerator, which renders the bounding ineffectual. Therefore, this experiment arm does not properly test bounded recall, though it does highlight the challenge of correct implementation in a limited-capacity setting.

Finally, feedback from the interviewers suggested 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. This finding is significant to the evaluation of the methods because often HIES take advantage of interviewer presence particularly in rural areas to collect additional information, such as prices, facility surveys, water testing, etc. Having more time to collect this auxiliary information boosts the analytical power of the data collection and enables new types of analysis without increasing costs.

4.2. Response rates

Bias in results can come from both unit and item non-response. Unit non-response refers to entire units of observation, in this case households, that do not participate in the survey either due to failure to contact or refusal to participate. The expected sample size for the two highly monitored diary arms was 72 households, with 216 households expected for each of the other three arms of the experiment. However, the achieved sample size, replacement rate, and effective sample size varied widely. Arm 1, the highly monitored paper diary, had the highest completion rate of its originally selected households, 86 percent, compared to the low monitored paper diary, which had a completion rate of only 66 percent. Across all five methods, replacements ended up constituting a relatively constant share between 17 and 20 percent of completed interviews. Depending on the characteristics of those refusals, replacing can actually compound bias by substituting outlier non-respondents with median replacements.

Item non-response refers to missing information for individual questions or sections within a questionnaire. Of the 716 interviewed households, over 11 percent reported no food consumption. This is especially common with low monitored diaries, where 18 percent of respondents recorded no transactions for food consumption.[2] The highly monitored CAPI diary had very low levels of missing consumption information at less than 2 percent, while the recall methods had about 10 percent of surveys being unusable due to missing information. The two-visit recall and one-visit recall methods, however, differed in that in one-visit recall, the consumption and non-consumption sections of the questionnaire were collected at the same time, while in the two-visit recall, the household sections were collected during the first visit, and then the interviewer returned about one week later to conduct the food consumption recall interview. In one-visit recall, if a household refused to participate, it was replaced. In the two-visit recall, if a household completed the main household questionnaire but refused to complete the consumption section when the interviewer returned, it was not replaced and counted as non-response. Despite this possible vulnerability in the two-visit recall method, however, substantial differences between completion and replacement rates were not observed, and the incidence of incomplete questionnaires was similar.

[2] 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 (Beegle et al, 2012).

The effective completion rate, or the number of completed and analyzable questionnaires as a share of the initial targeted sample size, is just over two-thirds (68.5 percent) for low monitored diaries, compared to the other four modules for which it ranges between 82 percent and 88 percent (Table 2). This finding has implications for costing and analysis as the households without consumption data cannot be used for all the analyses. Re-weighting of the remaining consumption data may mitigate some of the non-response bias, but only on observable characteristics and only for those which have high quality auxiliary data available.

4.2 Balance between treatment arms

As described above, there are two main ways in which the choice of data collection method can impact later analysis: participation and data quality. Different levels of participation based on the treatment arm can be seen if households with low levels of education or households with high demands for their time declined to fill in the expenditure diary, causing them to be dropped from the survey and therefore leading to differences in the population between arms. To determine if we have different populations across treatment arms, we undertook a series of balance tests, based on the same underlying population.

As a first check to understand if our samples for each arm were balanced, we compare basic demographic variables between arms. Due to the refusals and incomplete surveys, the analyzable samples were not evenly distributed across the arms by geography, and multivariate analysis shows the differences in demographic characteristics by arm hold even when controlling for location. Across three basic demographic characteristics between the five experimental arms (household size, percentage of households with female heads, and the age composition of households), the average household size tends to be smaller for recall (arms 4 and 5) over diary households, there were lower numbers of female headed households responding to the low monitored PAPI diary (arm 2), and the recall surveys tended to have higher numbers of working age adults (age 15 – 59) compared to those under 15 or 60 and older, but years of education of the household head were relatively consistent (Table 3). As noted above, the underlying samples were not evenly distributed across the arms by geography, and this inequality could be driving the results (Figure 3). Multivariate analysis, however, shows the above observations hold even when we control for the location of the household.

4.3. Number of items

In the absence of diary fatigue, the number of items and total expenditure should be relatively consistent across diary days, allowing for some variation due to weekly shopping patterns. To prevent bias from these patterns, the first day was staggered over days of the week. The two PAPI diaries, however, 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 (Figure 4). With the low monitored diary, the decline in the number of transactions was from 6.2 on day 1 to 3.0 on day 14. This finding was robust to controlling for day of the week and for location using econometric analysis (Figure 5). Combining the data from the highly monitored and low monitored diaries, the rate of decline was 3.4 percent fewer transactions recorded per day, the same rate of diary fatigue seen in the 2009/10 PNG HIES (Figure 1), resulting in 35 percent fewer transactions recorded, on average, in the low monitored diaries.

While the diary fatigue reduces measured household food acquisition, 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 reported larger starting food stocks than ending food stocks. The combination of these two patterns sees apparent destocking of food being equivalent to about 4 percent of total expenditure (Figure 6). For the low 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 lower 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 (acquisitions) and outflows into the household (where net destocking counts as an inflow). Also, even though the net destocking is offsetting to underreporting of transactions, it is distortionary. For example, in a case in which a household purchases one loaf of

bread each day but does not complete the diary for 7 days, it would appear that the household consumed only 7 loaves of bread during the 14 days instead of 14 loaves over the 14-day diary period. In the same way, if the household purchases 10 kg of rice at the start of the diary period and does not report 4 kg of rice stocks at the end of the diary period, it would appear the household consumed 10 kg of rice in the last 14 days instead of 6 kg. Incomplete diaries and misreporting of stocks distorts spending and nutritional analysis.

4.4. Consumption

We calculated a per capita consumption estimate that included four components: food and frequent non-food expenses (recorded in either diaries or reported by recall and annualizing from the 14-day or 7-day periods); infrequent expenses that were only obtained by recall, over either 7-days, 1-month, 3-months or 12-months (annualized); items such as alcohol and tobacco whose value of consumption was obtained from both diary and recall (annualized); and imputed rents and the use value of durables. The latter were based on calculations and should not vary by arm as the variables supporting the calculations are asked in identical ways for all households. Comparisons for this section were done using weights which only adjusted for non-response at the cluster level as these weights best represent what would be available to a typical analyst.

For food and frequent non-food, which is the most impacted by the variations in survey module design, the low monitored diary yields average consumption values that are 72.5 percent of what the highly monitored diaries yield and 68.0 percent of what single visit recall yields. The actual underestimate of food and frequent non-food consumption when using low monitored diaries is likely even greater as the value of food consumption derived from the diaries is inflated by upward bias from apparent destocking of food, particularly in the case of less monitoring where unreported expenditure is partially offset by destocking. Therefore, in diary surveys where food stocks are not being measured and food acquisition is used as a proxy for consumption, the likely understatement of actual consumption when using low monitored diaries would be even greater. Total food and non-food consumption was similar across the highly monitored diaries and the single visit recall, with values of $2,000 for the highly monitored PAPI diary, $2,103 for the highly monitored CAPI diary, and $2,187 for the single visit recall. The problematic two-visit recall had a total value of $1,699 and the low monitored PAPI diary had a value of $1,486.

There was also substantial variation in the result calculated from data collected on imputed rent and durable assets, despite being collected in an identical manner across the five arms. When standard household weights are applied, data collected via the low monitored diary ($582) and the highly monitored CAPI diary ($449) were significantly lower than data collected via the two recall methods ($738 and $642 for one-visit and two-visit recall respectively), as well as lower than the highly monitored PAPI diary ($736) but the small sample size and resulting wide confidence intervals for this arm mean the difference is not statistically significant (Figure 7). As there are no differences in collection method for this component, these mode effects can most readily be interpreted as different groups of respondents choosing to participate in the various arms. Intuitively, it could be expected that the highly monitored CAPI diary would have a lower value than the PAPI version as the PAPI version requires higher literacy levels than the CAPI diary. The low monitored diary similarly has fewer requirements because it is hypothesized that respondents simply did not complete it. The explanation for the higher values for the recall arms is less intuitive but could be interpreted as the differing opportunity costs of time. Since the recall sections are less time intensive, wealthier respondents may have higher refusal rates for diary keeping but not have differential response rates to less-wealthy households for a recall questionnaire, resulting in a composition effect of the responding sample by survey method.

To separate the impact of the mode effects (differences in responses rates between PAPI and CAPI leading to differences in the composition of respondents across survey arms) and questionnaire design effects (differences in responses from the way the question was asked), we calculate a second set of weights which use raking (Deville et al. 1993) to align a set of characteristics that we believe account for the differences in composition – average household size, literacy of the household head, share of female headed households, years of education of the household head, job category of the household head, and the calculated component of non-food consumption – leaving all remaining differences attributable to questionnaire design effects (Table 4). Since for the purposes of this experiment, we are concerned mainly with the comparisons between the arms rather than generalizing the results to the population of the Marshall Islands, we choose to align the values of the parameters in each arm to the overall averages across the five arms, even though those means are likely biased. This type of weighting would not be

possible in analysis where the effect itself is being studied, rather than the relationship between treatment arms, unless the value of the parameters in the overall populations were known. The raking methodology aligns more closely with the literature of official statistics than economics with the overall objectives being similar to those of regression or inverse-probability approaches, though raking does not require the linearity assumptions of the former and is more friendly to multi-arm experiments than the latter.

The values generated using the raking weights show only slight differences for the highly monitored PAPI, low monitored PAPI, and the two-visit recall, with less than a 5 percent change between the highest and lowest weighted values (Figure 8). The raking weights have the largest impact on the one-visit recall, reducing the estimated total per capita consumption from an unweighted value of $3,752 to a value of $3,495 with the raking weights, and for the highly monitored CAPI diary, increasing the estimated average from $3,161 to $3,492. The reweighted values for the one-visit recall and the highly monitored CAPI are then effectively identical, which indicates that once the respondent composition is netted out of the analysis, there is no difference in the results between the highly monitored CAPI diary and one-visit recall.

4.5. Food away from home

Food away from home represents an important component of food consumption contributing to around 17 percent of total food expenditure. The share of FAFH in the nationally representative 2019 HIES for Marshall Islands, which included an individually administered 7-day FAFH consumption recall module, was 20.5 percent of food expenditure and 15.2 percent of dietary energy consumption, respectively (Troubat and Sharp, 2021). The experiment found a substantial impact of survey design on the total FAFH consumption. While average per person per day consumption of FAFH of the high-monitored CAPI arm and one-visit 7-day FAFH consumption arm were comparable, values were markedly lower for the paper collection diary modes, with mean of FAFH recorded in the low-monitored paper diary arm two-thirds of that recorded in high-monitored CAPI arm (Figure 9). Due to the high variation and small sample sizes, however, these differences were not statistically significant, highlighting the need for further research.

4.6. Poverty measurement

Finally, we consider the implications of these measurement differences for poverty analysis using both national and international poverty lines. For the national line, we constructed a relative poverty line set as half the median real per capita consumption where real consumption uses spatial deflation based on a Fisher price index. These findings largely mirror those from per capita consumption. Using standard population weights, the highest poverty rate, 37.1 percent, is found with the low monitored diary, while there are similar rates (5.2 percent, 6.6 percent, and 8.1[3] percent) for the highly monitored PAPI diary, the highly monitored CAPI diary, and the one-visit recall, respectively (Figure 10). Perhaps more worrying, the different survey designs given differing profiles of the poor, which could then lead to erroneous policy conclusions based on the data. While all five methods show higher poverty in rural areas compared to the urban centers of Majuro and Kwajalein, the low monitored PAPI method and the two-visit recall show comparatively high poverty in Kwajalein, while the remaining three methods show low poverty levels in both Majuro and Kwajalein (Figure 11). The profiles of the poor are also distorted by survey design and differing response rates. For the highly monitored PAPI and the low monitored PAPI, there is no difference in the head's education between the poor and non-poor households, but the other three methods show significantly more education for non-poor heads. The low monitored PAPI diary shows the lowest poverty headcount for those households in which the head works in agriculture, while the one-visit recall shows the opposite (Figure 12). As the data from HIES surveys is an essential national policy making tool, distortions based on data collection method can lead to erroneous policy conclusions and misdirected or inefficient allocation of limited resources. With regard to the international extreme poverty line (IPL) of $1.90 per day, using the low monitored diary generates an IPL headcount rate of 8.9 percent, on-par with Vanuatu, one of the poorest countries in the region, compared with an IPL headcount rate of 4.2 percent with the highly monitored CAPI diary, on par with Fiji, one of the richest countries in the region.

4.7. Cost

As demonstrated above, both the highly monitored CAPI and the one-visit recall can generate high quality consumption data, and therefore the choice between the methods can be informed by the implementation costs. We calculated the cost of fielding each type of method in two ways. The first was based on what the survey budget would

[3] The basic needs poverty rate is 7.2 percent based on the 2019 household income and expenditure survey (World Bank, 2022).

be if the survey had exclusively used one method. The costs are calculated separately for rural and urban sectors because rural EAs had higher transport and labor 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 5). Survey costs in the Marshall Islands 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 labor costs, per-visit travel costs, printing, coding and data entry (Table 6). The two approaches yielded relative cost ratios that are similar, with highly monitored diaries costing almost five times as much as a one-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 low monitored diaries are at least twice as expensive as a one-visit recall, while a two-visit recall survey would be about 30 percent more costly than a one-visit survey. These cost estimates are all based on the completed households (n=716) rather than the households with analyzable data (n=632) 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).

In our experiment the *status quo* diary-keeping surveys have total costs that range from US$1,160 to US$3,020 per household in the rural sector (and $550 to $1,220 in urban areas). The range reflects whether revisit frequency was every two days or every week, which affects labor 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. The high costs as compared to surveys conducted in other contexts 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 one-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 similar to results from Tanzania (Beegle et al, 2012) which estimated 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 low 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.

5. Conclusions

The objective of this research was to understand the implications of the choice of survey design – both mode and consumption measurement methodology -- for poverty measurement in the Pacific Islands context. Our work is based on a randomized 5-arm experiment implemented in partnership with EPPSO, the national statistics office in the Marshall Islands, but the findings have wider implications for other small island and high data collection cost contexts. In the Pacific, the early recommendations from this paper have already led to changes in methodology for the 2019 rounds of HIES data collection in the Marshall Island, Kiribati, Vanuatu, and Wallis and Futuna.

The experiment demonstrated that response rates vary by arm, indicating the presence of mode effects. We found this at the unit level, for example, female headed households were less likely to participate in the low monitored diary and recall households had lower dependency ratios. Households that participated in the highly monitored PAPI and recall results were also better off as measured by the imputed values of housing and durable goods. This finding has two implications. First, the choice of method matters to respondent participation, which could either compound or offset errors, and introduce bias during the questionnaire completion process. Awareness of the groups who are less likely to participate in the survey can be integrated into training and supervision materials to decrease the impact of this source of bias. Secondly, the impact of the survey design demonstrates that switching methodologies breaks the poverty trend. Countries making this transition should either include a calibration experiment to understand the impact or refrain from making direct comparisons across surveys of different design.

Assuming respondent participation issues can be managed through improved training and supervision, the findings from this paper indicate that recall data collection is the recommended method to collect consumption data in the Pacific context. Using weights designed to compensate for differing respondents, the total per capita consumption amounts were near identical using the highly monitored CAPI diary and the one-visit recall, though the cost of the diary was nearly five times higher than recall. In addition, the experiment has clearly demonstrated evidence of diary fatigue in both the high monitored and low monitored diaries, with a decrease of nearly 50 percent in the number of transactions recorded on the first day compared to day 14 for the high monitored diaries, and an even more dramatic reduction over time for the low monitored diaries. Zero recorded consumption data also occurred nearly twice as often with paper diaries compared to recall methods. Data quality in diaries was further compromised by apparent net destocking that was most likely due to diminished cooperation by day 14, which upwardly biases consumption. While this offsets some of the error from diary fatigue, it is distortionary in terms of the basket and relative shares of items consumed, which has implications for poverty line construction as well as other uses of HIES data such as Consumer Price Index measurement and nutrition analysis.

Based on these findings, the 2019 round of HIES data collection in the Marshall Island, Kiribati and Vanuatu used the 7-day recall method. Also based on the results of this experiment, FAFH was collected with an individually administered module. The adoption of a dedicated FAFH module has increased the share of consumption away from home in three of the four recently completed surveys in the Pacific, with the exception of in Vanuatu where FAFH expenditure remained on-par with those surveys without a dedicated FAFH module (Figure 13; Sharp and Troubat, 2022).

The findings in this paper support those noted previously on the benefits of recall over diary data collection for food items. This paper, however, makes an additional contribution to the literature in that the data was collected by a national statistics office using their team of usual interviewers. In that way, the results are more typical of what could be expected in a real world situation as opposed to an academic experiment with more tightly controlled data quality protocols. In addition, these findings provide a basis for hope that regions like the Pacific, where some countries are far away from meeting the 2030 poverty goals, could develop a much more effective poverty monitoring

survey infrastructure based on more frequent and more timely data if they would move away from their tradition of relying on diary-keeping surveys.

References

Borlizzi, A., Delgrossi, M. E., & Cafiero, C. (2017). National food security assessment through the analysis of food consumption data from Household Consumption and Expenditure Surveys: The case of Brazil's Pesquisa de Orçamento Familiares 2008/09. Food policy, 72, 20-26.

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.

Claro, R. M., Baraldi, L. G., Martins, A. P. B., Bandoni, D. H., & Levy, R. B. (2014). Trends in spending on eating away from home in Brazil, 2002-2003 to 2008-2009. *Cadernos de saude publica*, 30, 1418-1426.

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

de Brauw, A., & Herskowitz, S. (2021). Income variability, evolving diets, and elasticity estimation of demand for processed foods in Nigeria. *American Journal of Agricultural Economics*, *103*(4), 1294-1313.

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.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.

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.

Farfán, G., Genoni, M. E., & Vakis, R. (2017). You are what (and where) you eat: capturing food away from home in welfare measures. *Food Policy*, *72*, 146-156.

FAO and The World Bank. 2018. Food Data Collection in Household Consumption and Expenditure Surveys. Guidelines for Low- and Middle-Income Countries. Rome. 104 pp. Licence: CC BY-NC-SA 3.0 IGO.

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.

Dex, S., 1995. The reliability of recall data: A literature review. *Bulletin of Sociological Methodology / Bulletin de Methodologie Sociologique*, 49(1), pp.58-89.

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, and 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

Sharp, M, and Troubat, N. 2022. Experience in measurement of food consumption away from home in the Pacific region. Open Meeting of the United Nations Committee of Experts on Food Security, Agriculture and Rural Statistics (UN-CEAG). 17 February 2022.

Smith, L. C., Dupriez, O., & Troubat, N. (2014). Assessment of the reliability and relevance of the food data collected in national household consumption and expenditure surveys. *International Household Survey Network*.

Troubat, N. and Sharp, M.K. 2021. Food consumption in the Marshall Islands – Based on analysis of the 2019/20 Household Income and Expenditure Survey. Majuro, FAO and SPC. https://doi.org/10.4060/cb7583en.

Sudman S. and Bradburn N. 1974. Response effects in surveys: A review and synthesis. Adline, Chicago, IL.

Weisberg, H. F. 2005. The Total Survey Error Approach. Chicago: University of Chicago Press.

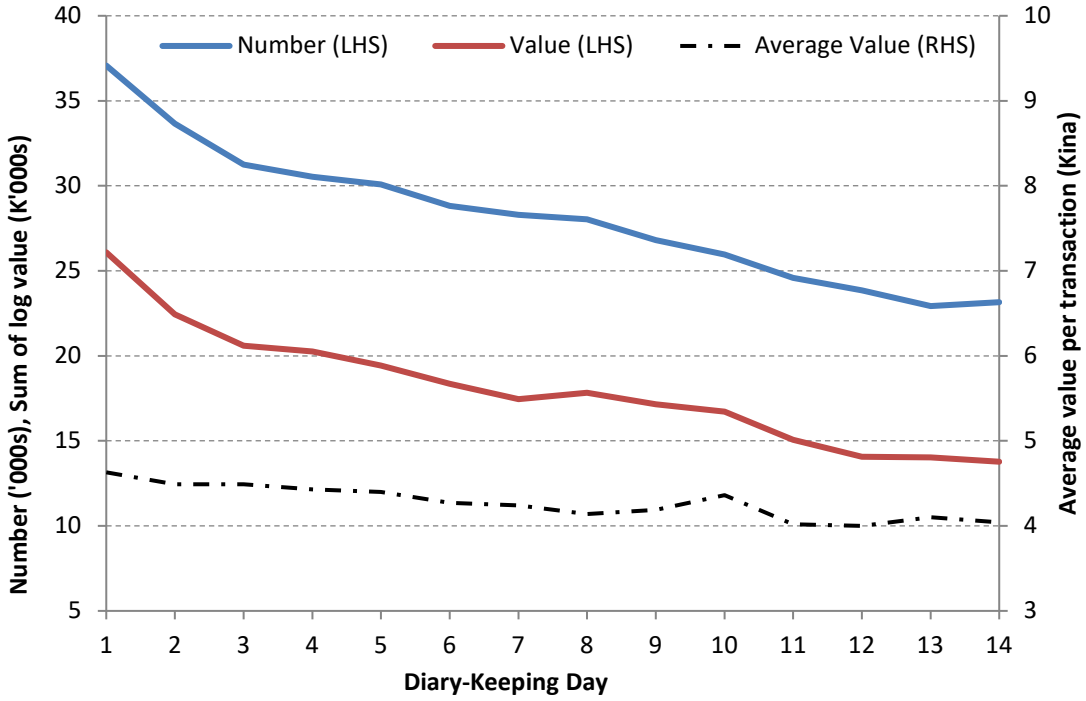
World Bank, 2018. Survey Solutions CAPI/CAWI platform: Release 5.26. Washington DC, The World Bank.

World Bank. (2022). Pacific Poverty Assessments Report 2021. World Bank Publications.

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.

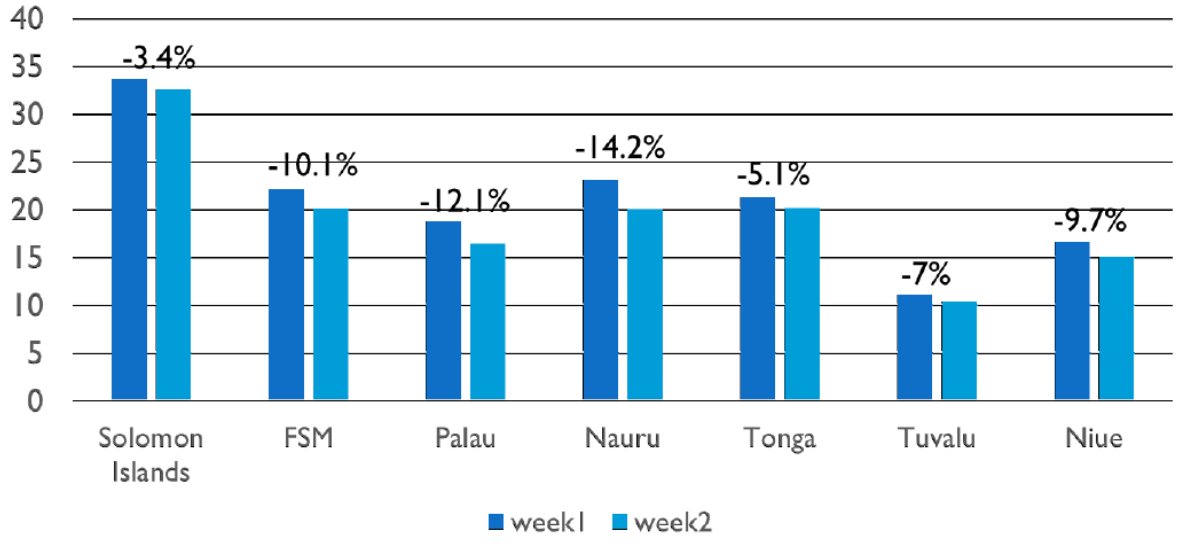
Figures

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



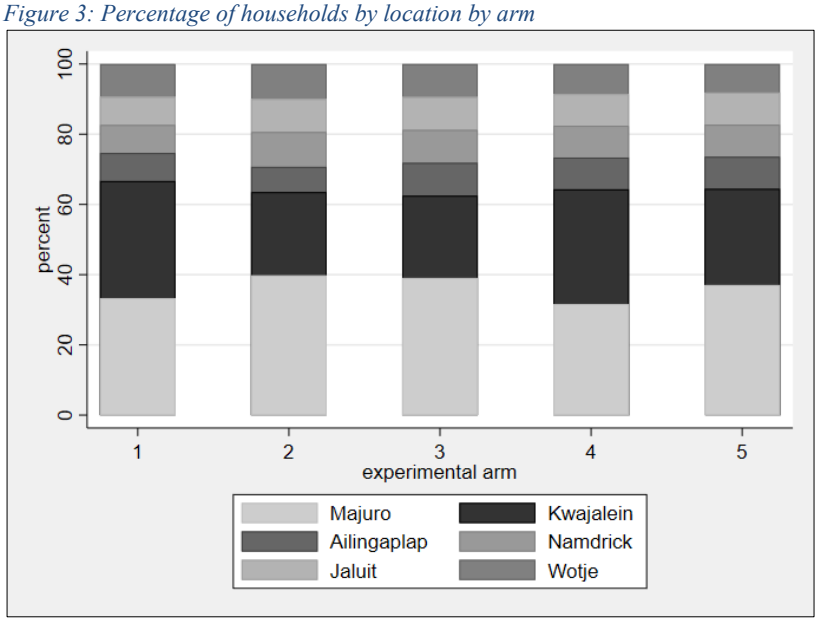
Source: Gibson 2013

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



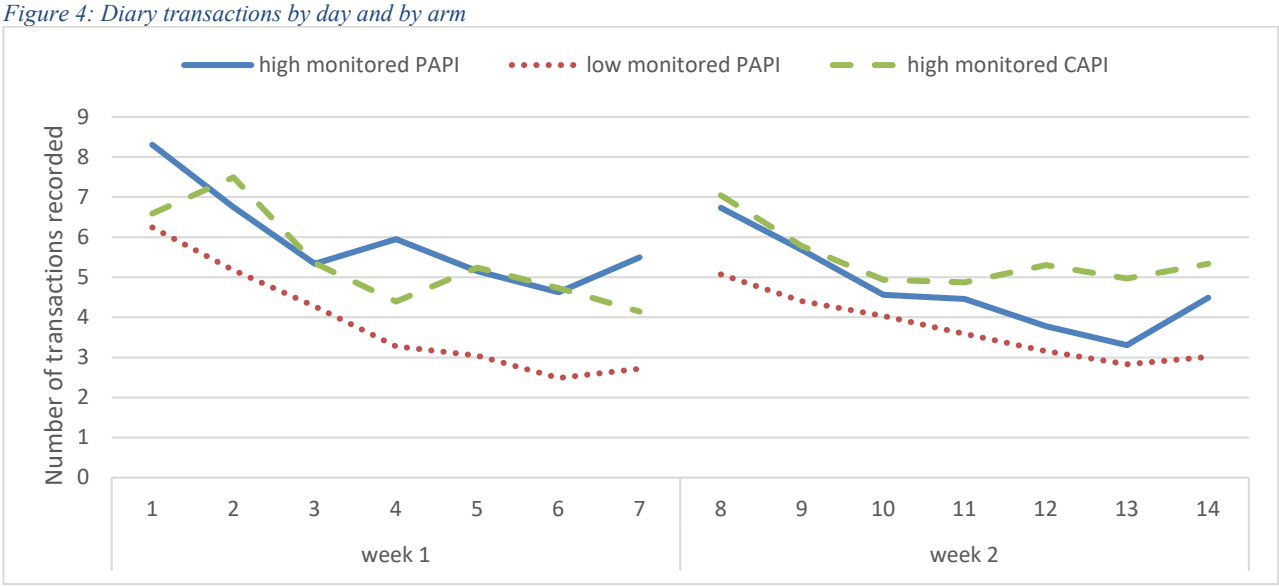
Source: Sharp et al. 2018

Figure 3: Percentage of households by location by arm



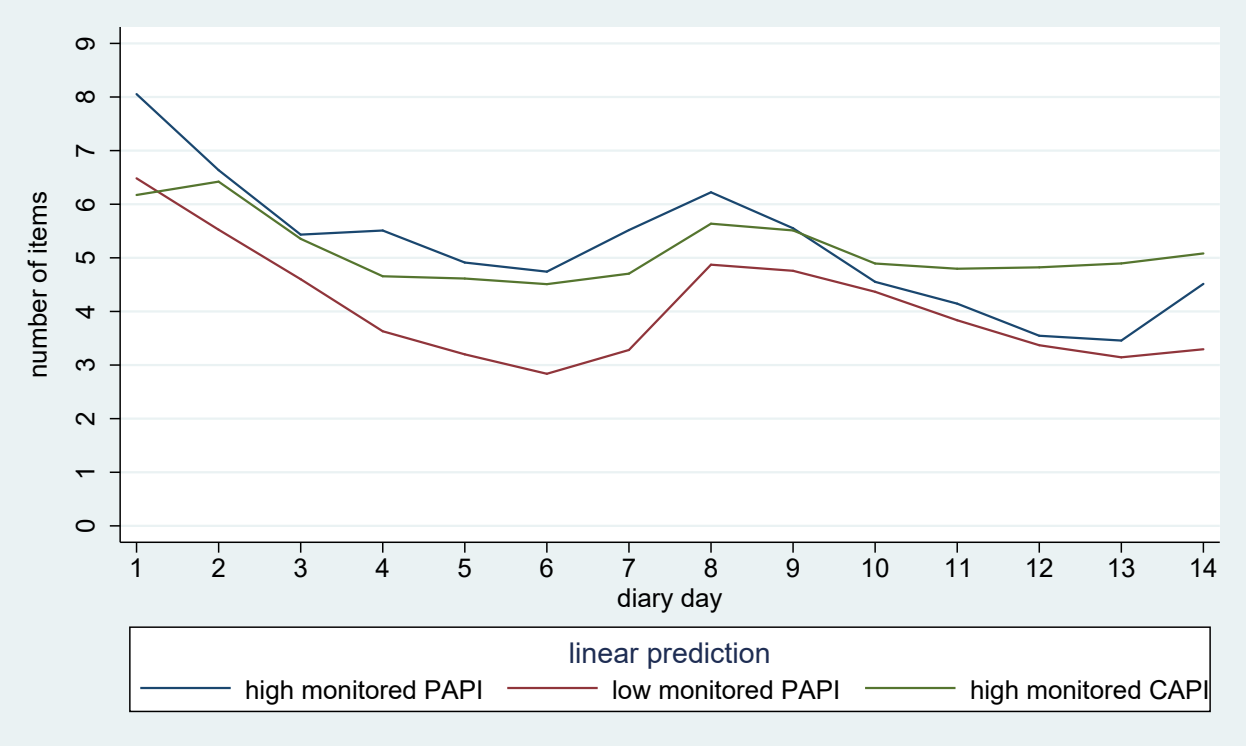
Source: Authors' calculations

Figure 4: Diary transactions by day and by arm



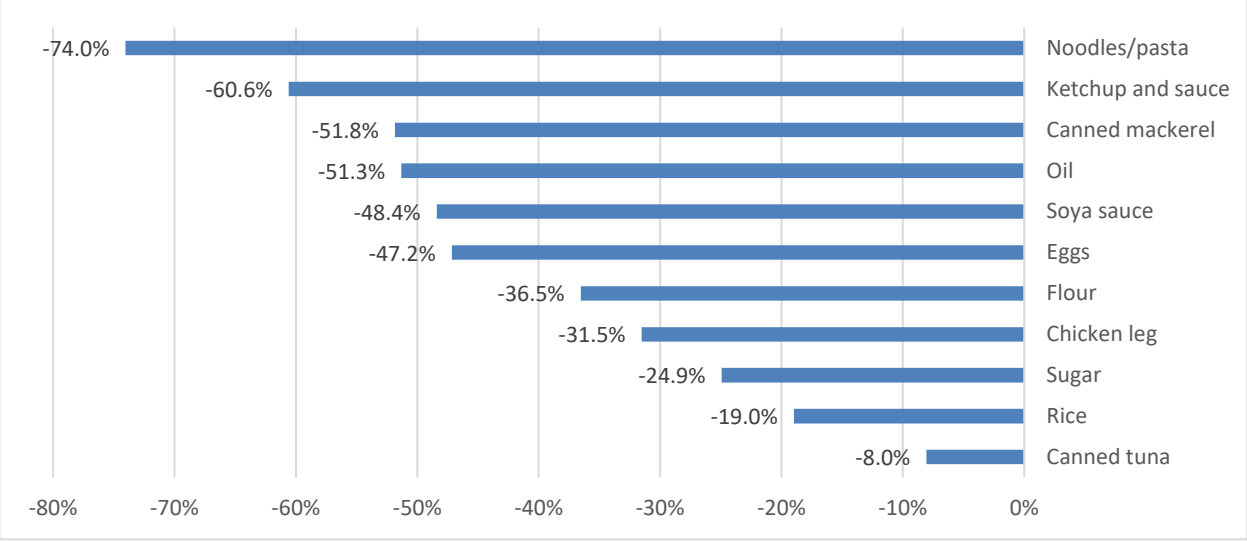
Source: Authors' calculations

Figure 5: Linear prediction for number of diary items (controlling for atoll and day of the week)



Source: Authors' calculations

Figure 6: Net destocking by item



Source: Authors' calculations

Figure 7. Components of consumption aggregate by collection method (standard household weights)

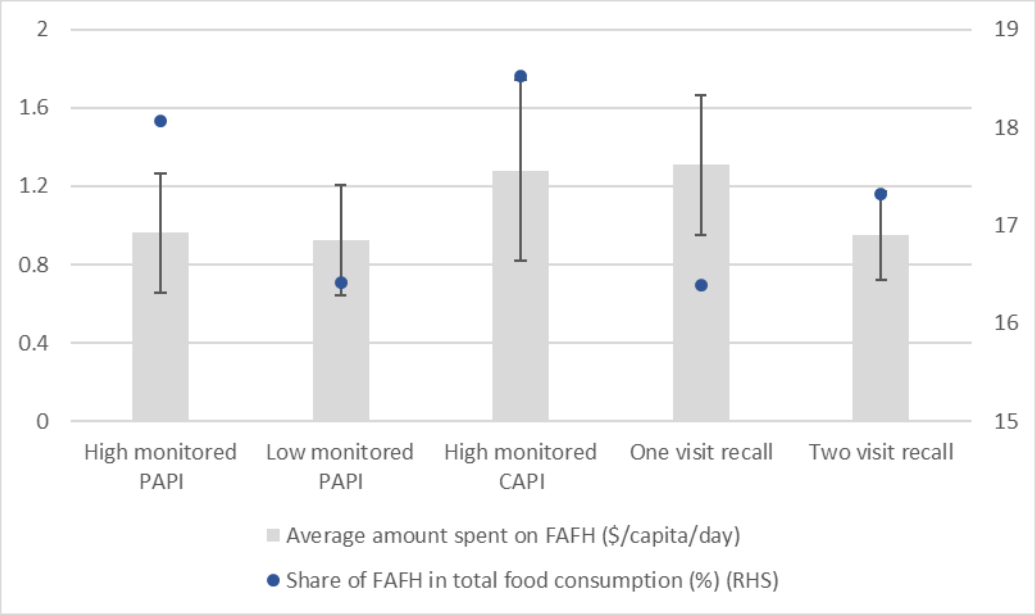
Source: Authors' calculations

Figure 8. Components of consumption aggregate by collection method (raking weights)

Source: Authors' calculations

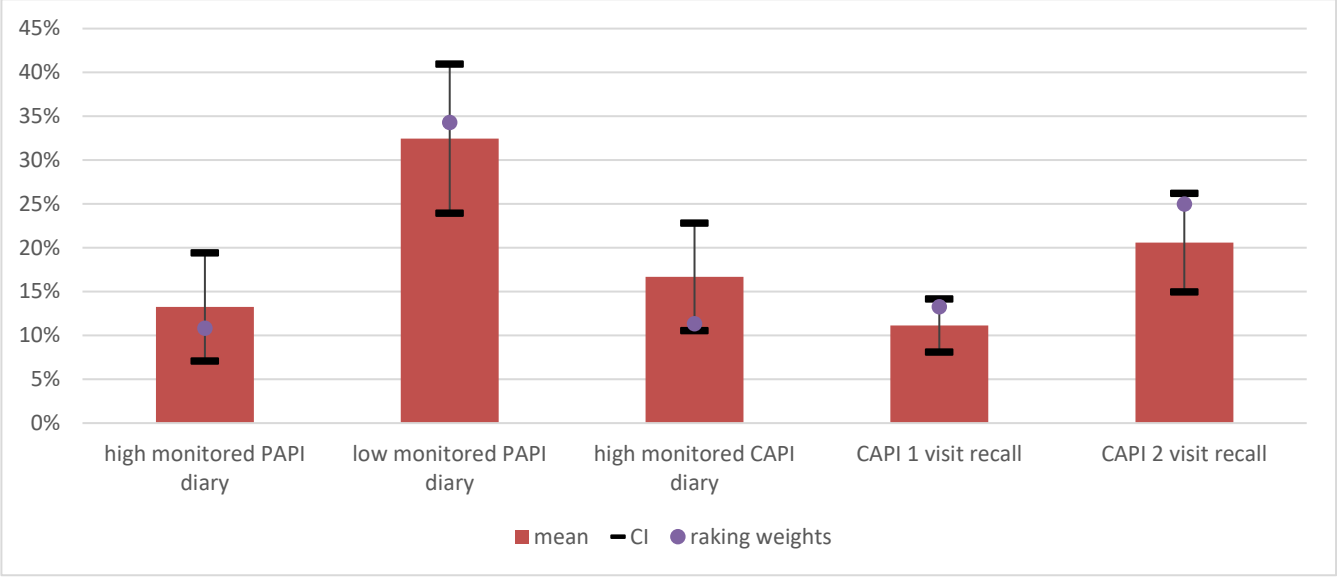
Note: The figures above show the total expenditure by method of calculations (x-axis) across the five experimental arms (y-axis) using two different weighting methods. Food and other frequently purchased expenditures – i.e. those captured with a diary in the diary-based arms and with one-week recall in the recall arms – are presented in the first line labeled "diary/recall." The second line labeled "recall" include infrequent expenditures, such as clothing, education, holiday travel, etc. that is capture using 3-month or one-year recall periods for both methods. The third line labeled "both" includes expenditures that can appear in both – such as over-the-counter medications and cell phone credit – depending on the respondent. "Calculated" values are the actual or imputed rent of the respondent's dwelling and the use-value of durable goods, which are calculated in the same way from the same set of questions regardless of experimental arm. The final line is the total of the four groupings.

Figure 9: Mean consumption of food away from home (USD/capita/day using ranked weights for the confidence intervals) and share of FAFH in total food consumption



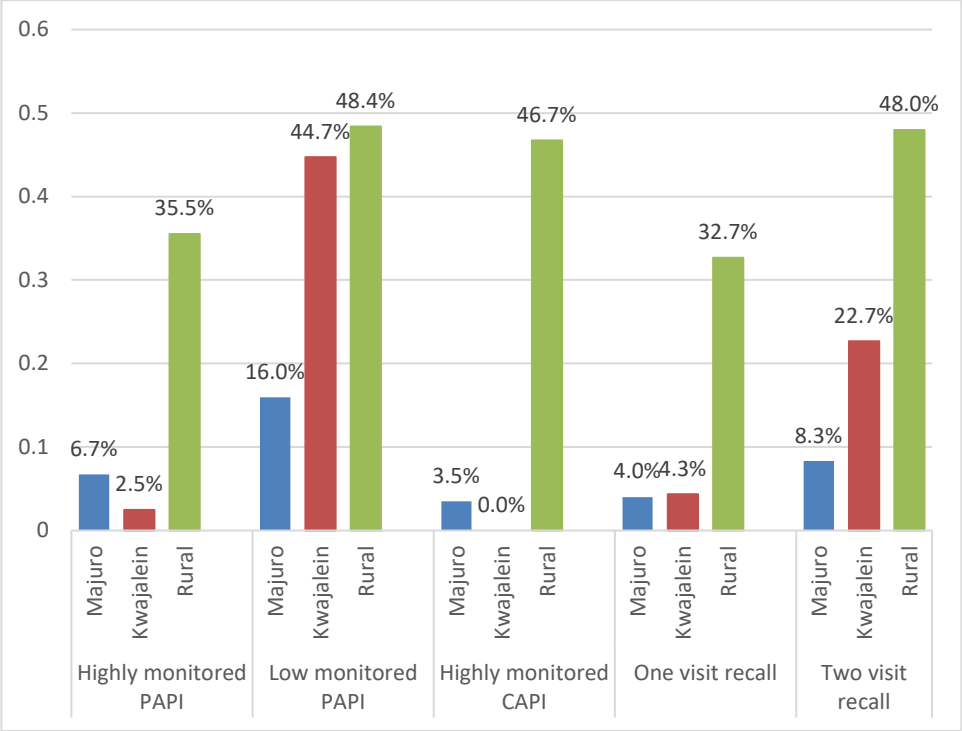
Source: Authors' calculations

Figure 10: Poverty Rates by Experiment Arm (Poverty Line defines as half of median)



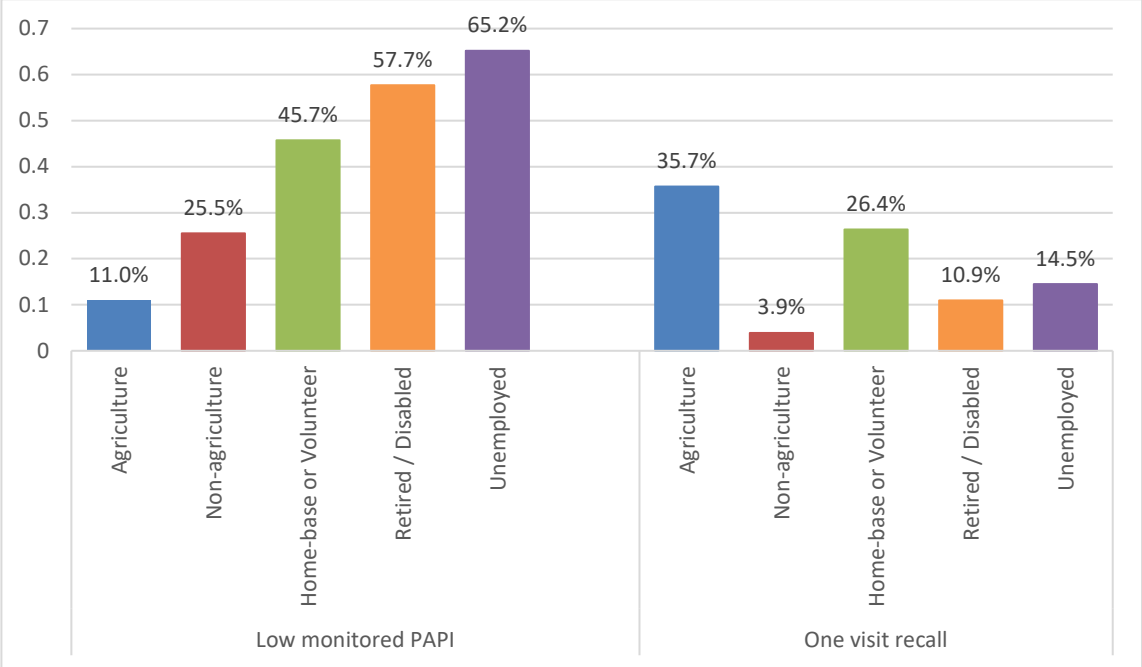
Source: Authors' calculations

Figure 11: Poverty headcount by location



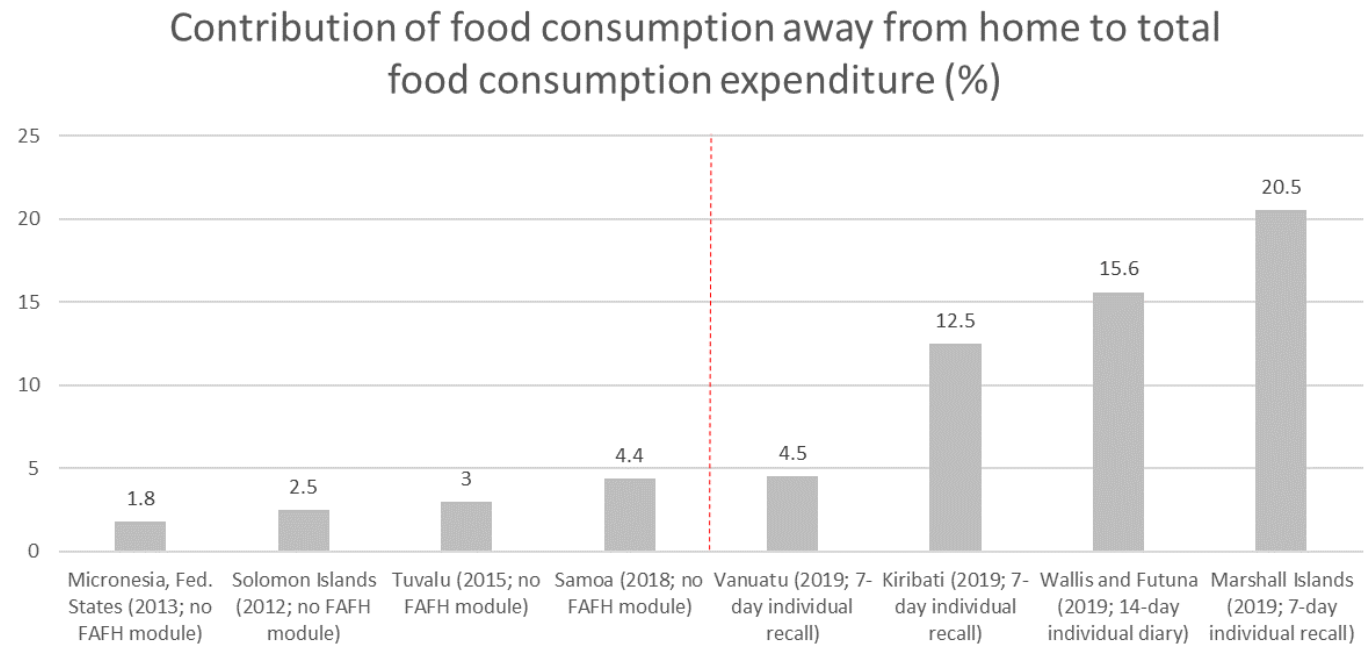
Source: Authors' calculations

Figure 12: Poverty headcount by employment sector of household head



Source: Authors' calculations

Figure 13: Contribution of food consumption away from home to total food consumption expenditure (%)



Note: Red line denotes old method to the left (no FAFH module) and new method (with individual FAFH module) to the right
Source: Sharp and Troubat, 2022

Tables

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			

Table 2. Description of Treatment Arms and Sample Sizes

		Design		Completion rate		Replacement rate				Partial interviews			Effective completion rate	
Arm	Description	Workload per interviewer	Target sample size	Complete interviews from original selection	Completion Rate	Complete interviews from replace-ment list	Replace-ments of unknown source	Collected sample size	Replace-ments as share of collected interviews	No reported food cons	Incomplete for other reason	Incomplete as share of collected interviews	Analyzable records	Effective completion rate
1	14-day diary, highly monitored (visits every 2 days), transactions recorded using pen and paper	6	72	62	86.11%	12	1	75	17.33%	16	0	21.33%	59	81.94%
2	14-day diary, less monitored (visits each week), transactions recorded using pen and paper	18	216	143	66.20%	33	3	181	19.89%	33	2	19.34%	148	68.52%
3	14-day diary, highly monitored (visits every 2 days), data entered by interviewer using CAPI during each visit	6	72	52	72.22%	12	0	64	18.75%	1	0	1.56%	63	87.50%
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	162	75.00%	35	2	199	18.59%	19	0	9.55%	180	83.33%
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	152	70.37%	36	2	197	19.29%	13	7	10.15%	184	85.19%
Total			792	571		128	8	716		82	9		634	80.05%

Table 3: Differences in demographic characteristics by arm

	household head characteristics			household composition		
arm	household size	female head	years of education	share 0-14	share 15-60	share 60+
1	5.54	40.68	10.17	31.25	54.63	14.11
2	5.51 *	25.00 **	10.75	29.19	58.22	12.59
3	5.73	30.16	10.62	31.22	60.08	8.69
4	4.59 **	33.33	10.91	27.11	62.20	10.69
5	4.87	29.67	10.74	25.24 *	64.47 **	10.29
experiment mean	5.15	32.26	10.75	27.99	60.84	11.17

Note: Indicates statistical difference from experiment mean (10%, ** 5%, *** 1%)*

	employment of household head					
arm	agriculture	outside agriculture	unemployed	home-based or volunteer	retired or disabled	study
1	15.25	28.81	10.17	33.90	11.86	0.00
2	6.76	40.54	5.41	40.54	6.76	0.00
3	7.94	49.21	3.17	31.75	7.94	0.00
4	12.78	48.89	7.22	20.00	10.00	1.11
5	15.38	34.07	11.54	24.18	13.19	1.65
experiment mean	11.87	40.82	7.91	28.48	10.13	0.79

Pearson chi2(20) = 43.2341 Pr = 0.002

Table 4: Illustrating the impact of the raking weights for selected demographic variables

		Unweighted sample						Weighted sample					
		Arm 1	Arm 2	Arm 3	Arm 4	Arm 5	experimental mean	Arm 1	Arm 2	Arm 3	Arm 4	Arm 5	experimental mean
		Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE	Mean/SE
	Household size	5.54	5.51*	5.73	4.59***	4.87	5.09	5.15	5.15	5.15	5.15	5.15	5.15
		[0.366]	*[0.192]*	*[0.384]*	*[0.187]*	*[0.200]*	*[0.105]*	*[0.355]*	*[0.189]*	*[0.357]*	*[0.223]*	*[0.218]*	*[0.110]*
Household head characteristics	Female head	40.68	25.00	30.16	33.33	29.67	30.70	32.26	32.26	32.26	32.26	32.26	32.26
		[6.450]	*[3.571]*	*[5.829]*	*[3.523]*	*[3.395]*	*[1.836]*	*[6.138]*	*[3.856]*	*[5.937]*	*[3.494]*	*[3.475]*	*[1.861]*
	Years of education	10.17	10.75	10.62	10.91	10.74	10.73	10.75	10.75	10.75	10.75	10.75	10.75
		[0.379]	*[0.202]*	*[0.354]*	*[0.268]*	*[0.237]*	*[0.123]*	*[0.370]*	*[0.201]*	*[0.370]*	*[0.255]*	*[0.239]*	*[0.119]*
	Share 0-14[(1)]	31.25	29.19	31.22	27.11	25.24*	27.86	29.75	26.14	30.22	27.94	25.90	27.99
		[2.299]	*[1.661]*	*[2.624]*	*[1.636]*	*[1.592]*	*[0.008]*	*[2.356]*	*[1.613]*	*[2.687]*	*[1.601]*	*[1.560]*	*[0.809]*
Household composition	Share 15-60[(1)]	54.63**	58.22	60.08	62.20	64.47**	61.00	57.30	59.79	60.98	61.22	64.93	60.84
		[2.951]	*[1.828]*	*[3.093]*	*[1.878]*	*[1.805]*	*[0.010]*	*[2.948]*	*[1.932]*	*[3.221]*	*[1.801]*	*[1.751]*	*[0.952]*
	Share 60+[(1)]	14.11	12.59	8.69	10.69	10.29	11.14	12.95	14.07	8.80	10.85	9.17	11.17
		[2.909]	*[1.627]*	*[2.262]*	*[1.428]*	*[1.453]*	*[0.008]*	*[2.868]*	*[1.616]*	*[2.419]*	*[1.350]*	*[1.341]*	*[0.775]*
	Agriculture	15.25	6.76	7.94	12.78	15.38**	11.87	10.07	10.07	10.07	10.07	10.07	10.07
		[4.721]	*[2.070]*	*[3.433]*	*[2.495]*	*[2.682]*	*[1.287]*	*[3.952]*	*[2.482]*	*[3.822]*	*[2.250]*	*[2.237]*	*[1.198]*
	Outside agriculture	28.81**	40.54	49.21	48.89*	34.07**	40.82	42.35	42.35	42.35	42.35	42.35	42.35
		[5.947]	*[4.049]*	*[6.349]*	*[3.736]*	*[3.523]*	*[1.957]*	*[6.488]*	*[4.075]*	*[6.275]*	*[3.693]*	*[3.673]*	*[1.967]*
	Unemployed	10.17	5.41	3.17**	7.22	11.54	7.91	7.98	7.98	7.98	7.98	7.98	7.98
		[3.969]	*[1.865]*	*[2.227]*	*[1.935]*	*[2.375]*	*[1.075]*	*[3.558]*	*[2.235]*	*[3.442]*	*[2.026]*	*[2.014]*	*[1.079]*
Employment of household head	Home based or volunteer	33.90	40.54***	31.75	20.00***	24.18	28.48	28.93	28.93	28.93	28.93	28.93	28.93
		[6.216]	*[4.049]*	*[5.912]*	*[2.990]*	*[3.182]*	*[1.797]*	*[5.954]*	*[3.740]*	*[5.759]*	*[3.389]*	*[3.370]*	*[1.805]*
	Retired or disabled[(1)]	11.86	6.76*	7.94	10.00	13.19	10.13	10.67	10.67	10.67	9.91	9.57	10.30
		[4.246]	*[2.070]*	*[3.433]*	*[2.242]*	*[2.515]*	*[1.201]*	*[4.054]*	*[2.546]*	*[3.921]*	*[2.234]*	*[2.187]*	*[1.210]*
	Study[(1)]	0.00	0.00	0.00	1.11	1.65	0.79	0.00	0.00	0.00	0.76	1.10	0.37
		[0.000]	*[0.000]*	*[0.000]*	*[0.783]*	*[0.946]*	*[0.353]*	*[0.000]*	*[0.000]*	*[0.000]*	*[0.648]*	*[0.775]*	*[0.242]*
	N	59	148	63	180	182	632[(2)]	59	148	63	180	182	632

The value displayed for t-tests are the differences between the unweighted mean and the weighted mean for each arm.
***, **, and * indicate significance at the 1, 5, and 10 percent critical level.
(1) Not used to rake the weights
(2) Two outliers removed

Table 5: 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.

Table 6: 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