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

Measuring Poverty Rapidly Using Statistical Imputations

Utz Pape

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

Measuring Poverty Rapidly Using Statistical Imputations

Utz Pape[1]

Abstract

Poverty is an indicator of paramount importance for gauging the socioeconomic wellbeing of a population. Especially during or after a shock, poverty estimates are invaluable for assessing the severity of the impact and for identifying which parts of the population were most affected. The measurement of consumption, however, has traditionally been very time consuming. A household consumption questionnaire usually includes more than 200 items, including both food and nonfood items, often requiring more than two hours to administer. This paper proposes a new methodology that combines an innovative questionnaire design with standard imputation techniques. It substantially shortens the time required to administer a household consumption questionnaire to less than 60 minutes by imputing deliberately absent consumption values for items that are not explicitly asked. The proposed methodology makes it possible to derive poverty estimates without compromising the credibility of the resulting estimate, and it performs considerably better than alternative approaches based on reduced consumption aggregates and cross-survey imputations. This new methodology is particularly useful in fragile states given the significant risks associated with lengthy interviews. It can also be useful to reduce enumerator and respondent fatigue, or to mitigate the problem of high non-response rates.

Keywords: Poverty and inequality measurement, survey methods

JEL: C83, D63, I32

[1] Corresponding author: Utz Pape (upape@worldbank.org). The findings, interpretations and conclusions expressed in this paper are entirely those of the authors, and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments of the countries they represent. The authors would like to thank Chris Elbers for help with the statistical properties of the new methodology, and Kathleen Beegle, Tomoki Fujii, Kristen Himelein, Dean Jolliffe, Peter Lanjouw, Emmanuel Skoufias, Shinya Takamatsu, Roy Van der Weide and Nobuo Yoshida for discussions, as well as Johan Mistiaen for discussions and support in pursuing the idea.

Introduction

Poverty is an indicator of paramount importance for gauging the socioeconomic wellbeing of a population. Especially during or after a shock, poverty estimates are invaluable for understanding the situation, as well as for assessing the severity of the impact and for identifying which parts of the population were most affected. Especially in the developing world, consumption-based poverty measures are used, defining the poor as those households with consumption levels that fall below a set poverty line (Deaton and Zaidi 2002). The poverty line is usually set at a consumption level adequate for sustaining the minimum level of welfare required for healthy living (Ravallion 1998). Consumption-based poverty measures are widely used in development contexts and play a critical role in policy decisions (e.g. Beegle et al. 2016).

The measurement of consumption, however, has traditionally been very time consuming. A typical household consumption questionnaire contains a series of questions about the price and quantity consumed for each item, and whether it has been purchased, self-produced, or bartered. Usually encompassing more than 200 food and nonfood items, the time required to administer such a questionnaire can often substantially exceed two hours. In addition to high administration costs due to long interview times, measurement errors may become significant towards the end of the questionnaire as enumerators and respondents become fatigued. Respondents might also cancel the interview before it is completed, thus contributing to a higher non-response rate.

To overcome the challenges inherent to measuring consumption poverty, we propose a new methodology that combines an innovative questionnaire design with standard imputation techniques.[2] This new methodology allows us to substantially shorten the consumption questionnaire and reduce the interview time (less than 60 minutes for a standard questionnaire) by imputing deliberately absent consumption values for those items that are not explicitly asked about. Poverty estimates can be derived in this way without compromising the credibility of the resulting estimate. This new methodology is particularly useful in fragile states given the significant risks associated with lengthy interviews. It can also be useful to reduce enumerator and respondent fatigue, or to mitigate the problem of high non-response rates.

The most straightforward way to reduce the expected interview time is to skip rarely-consumed items. Another simple strategy is to ask the respondent about an aggregate amount of spending on an entire category of consumption (e.g., total expenditure on flour) instead of individual items (e.g., expenditure on corn flour, wheat flour, etc.). However, altering the set of items in the questionnaire can result in a nontrivial change in the reported consumption amount (Olson-Lanjouw and Lanjouw 2001). Both approaches are likely to lead to an underestimation of consumption and overestimation of poverty, as was demonstrated in a study in Tanzania that directly compared various methods of measuring consumption (Beegle et al. 2012).

An alternative approach is to apply methods of cross-survey imputation. In situations where full household expenditure surveys are too costly or impractical to administer, data gaps can be filled using other surveys that have common covariates that are correlated with household expenditure. For example,

---

[2] A precursor methodology based on the same principle was previously published in Pape and Mistiaen 2015.

data from a full consumption survey can be combined with data from shorter and more frequent labor force surveys to generate poverty estimates (Douidich et al. 2013). While such methods may work well even when there is a rapid economic change (Christiaensen et al. 2011), the assumption of a stable structural parameter typically cannot be tested and may not be valid, especially in the context of large and systemic shocks or if a substantial amount of time has passed since the baseline survey was implemented. It is also possible to design a survey such that one sample has a full consumption module and another sample has only the covariates of consumption. Consumption can thus be imputed and poverty estimates can be derived at a reduced cost, even though the magnitude of potential cost reduction may be modest (Fujii and van der Weide 2016). In such a setup, however, the sample for the full consumption module must be chosen randomly to avoid biased estimates of the model parameters. Thus, this approach is only of limited usability in the case of fragile countries as it might not be feasible administering the full consumption module in particular insecure areas, creating a downward bias in poverty estimates for those areas.

This paper is organized as follows. We first present the proposed methodology with its statistical properties as well as the data in Section 2. In Section 3, we apply the methodology to different scenarios showing the tradeoffs between performance and parameters of the approach, and then compare it to a reduced consumption approach as well as a more sophisticated reduced consumption approach adjusting the poverty line. The section ends with a real-world example based on a pilot survey in Kenya, assessing the performance of the new methodology and comparing it to a cross-survey imputation approach. The paper finishes with Section 4 concluding the findings and discussing some of the limitations of the new approach.

Methodology

Overview

The rapid approach being proposed here applies a split-questionnaire design to the consumption module of a household survey, thereby generating systematically absent data that can be conveniently imputed.[3] Instead of having all households report on all consumption items, important items are assigned to a core module and the remaining items are split into two or more optional modules.[4] Each household then answers the questions in the core module and in only one of the optional modules. This approach reduces average interview time considerably, down to 45 to 60 minutes per household for a standard household consumption survey. The cost of this efficiency gain is that data are deliberately absent for those optional modules that were not administered to certain households. We can however offset this cost by estimating the deliberately absent data for each household based on the data collected from other households for that module. While this approach utilizes a structural model for the imputation of the deliberately absent data, the model is estimated within the survey rather than between two surveys, thereby circumventing

[3] While the split-questionnaire design is more popular in other disciplines such as psychology (e.g. Graham, Hofer, and MacKinnon 1996), the approach has not yet been applied to large-scale household-based surveys, nor with the goal of reducing the time required to estimate consumption or poverty.

[4] As is shown below, the core module is not strictly necessary further reducing the interview time.

the problem of biased structural parameters due to having different sample populations or considering the same population at different points in time.

The rapid approach starts by defining the number of core items and the number of optional modules for the non-core consumption items. The smaller the number of core items and the greater the number of optional modules used, the faster the questionnaire can be administered, as fewer items need to be asked for each household. However, having less core items and more optional modules also increases the uncertainty in the estimation as less consumption information is available. Thus, the choice of these two parameters can be informed by simulations on a previous or similar survey to gauge the performance of the estimation vis-à-vis the time savings in administering the questionnaire. Another consideration is that it is beneficial to balance the number of households for each optional module, ideally at the cluster level of the survey.

The next step is to select core consumption items. Although consumption in any given country will exhibit some variability, data on a few dozen key items will usually be sufficient to capture the majority of consumption. Important consumption items can be identified using average consumption share per household or across households, as estimated by previous consumption surveys in the same context or recorded consumption shares in neighboring and/or similar countries. While a good choice of core items will improve the performance of the estimation, the methodology still works if no core items are used, e.g. in a context without any prior information. The identified key items are then assigned to the core module that will be administered to all households.

Finally, non-core items are randomly partitioned into optional modules. It is important to note that the conceptual distinction between core and optional items should not be reflected in the layout of the questionnaire. Instead, all items per household need to be grouped into categories of consumption items (e.g. meat, fruits, vegetables, cereals) and different recall periods. It is therefore recommended to use CAPI (Computer-Assisted Personal Interviewing) technology, which makes it possible to hide the modular structure of the consumption questions within the layout of the questionnaire.

Once the core and optional modules have been defined and the design has been finalized, the survey can be implemented. The assignment of optional modules to households is performed randomly and is stratified by enumeration area, thus ensuring an appropriate representation of all optional modules in each enumeration area. Once the data have been collected and cleaned, household consumption is estimated by imputation. The average consumption of each optional module can be estimated based on the sub-sample of households assigned to that optional module.

Theoretical Properties

Consumption for a household *i* is the sum of the consumption for each item in each module

$$y_i = \sum_k y_{ik} = \sum_k \sum_j y_{ikj}$$

where y_{ikj} denotes the consumption of item j in module k.[5] Applying the rapid approach, we only observe a subset of modules y_{ik}, specifically for each household k=0 and one other module where k>0. We can formalize this by using a binary (0,1) variable b_k, which is independent of y_{ik}, where $P(b_k=1) = \pi_k$. In practice, the assignment of optional modules can be done more systematically to ensure a balanced design at the cluster-level, which does not invalidate the assumed independence of b_k from consumption. The expected consumption of a household is:

$$Ey_i = E\sum_k y_{ik}\frac{b_k}{\pi_k} = \sum_k Ey_{ik}$$

We obtain a consistent and unbiased estimator for expected consumption if we can find consistent and unbiased estimators for expected module consumption. This also holds for regressions assuming b_k and household characteristics x_i are independent:

$$E(y_i|x_i) = E(\sum_k y_{ik}\frac{b_k}{\pi_k}|x_i) = \sum_k E(y_k|x_i)$$

Furthermore, the second moment can be estimated as follows:

$$Ey_i{}^2 = E(\sum_k y_{ik})^2 = E\sum_k y_{ik}^2 + 2E\sum_{k\neq l} y_{ik}y_{il} = \sum_k \frac{b_k}{\pi_k}Ey_{ik}^2 + 2\sum_{k\neq l}\frac{b_k}{\pi_k}\frac{b_l}{\pi_l}Ey_{ik}y_{il}$$

Similarly, higher moments can be constructed. Thus, the complete distributional information of y can theoretically be recovered from sufficiently large samples if the design of the split questionnaire allows for the estimation of correlations between modules.

Consumption Estimator

Distinguishing between administered core module $k = 0$, the administered optional module k_i^* and the non-administered remaining optional modules $0 < k \neq k_i^*$, we obtain as estimator for consumption

$$\hat{y}_i = y_{i0} + y_{ik_i^*} + \textstyle\sum_{k\neq k_i^*}\hat{y}_{ik}.$$

As shown above, the estimator is unbiased for Ey_i as

$$Ey_i = \sum_k Ey_{ik} = Ey_{i0} + Ey_{ik_i^*} + \sum_{k\neq k_i^*} E\hat{y}_{ik} = E\hat{y}_i$$

The variance of consumption can be decomposed as

$$Var(y_i) = Var(y_{i0}) + \sum_k Var(y_{ik}) + 2\sum_k Cov(y_{i0}, y_{ik}) + \sum_{k\neq l} Cov(y_{ik}, y_{il})$$

[5] Note that we assume consumption to be per-capita throughout the paper.

$$\geq Var(y_{i0}) + \sum_k Var(\hat{y}_{ik}) + 2\sum_k Cov(y_{i0}, \hat{y}_{ik}) = Var(\hat{y}_i)$$

with the inequality given by the assumption of positive correlation between optional modules.[6] The variance is thus underestimated, as we cannot measure correlation between modules and so assume them to be independent $Cov(y_{ik}, y_{il}) = 0$ for all optional modules *k* and *l*. The more optional modules are used, the higher the under-estimation of the variance. Contrarily, the larger the fraction of the variance captured in the core module, the lower the underestimation of the variance. This suggests using a low number of optional modules with a large number of items in the core module. This represents the fundamental tradeoff between the accuracy of the estimator and time savings, which are higher with more optional modules and fewer items in the core module.

We apply the Foster–Greer–Thorbecke measures of poverty (Foster, Greer, and Thorbecke 1984) to the consumption aggregate defined as

$$FGT_{\alpha,z} = N^{-1} \sum_{i:y_i<z} \left(\frac{z - y_i}{z}\right)^\alpha$$

where N denotes the number of households, z is the poverty line, and y_i is consumption for a given household. By selecting the coefficient α we can produce different poverty measures: $\alpha = 0$ for the poverty headcount, $\alpha = 1$ for poverty depth and $\alpha = 2$ for poverty severity. Given that we are underestimating the variance of y_i, this implies that the estimator $\widehat{FGT}_{\alpha,z}$ will be underestimated for a poverty line *z* smaller than the mode of y and overestimated for larger poverty lines.

Estimation

The optional module consumption can be estimated in the log-space conditional on strictly positive consumption:

$$\log y_{ik} = \beta X_i + \varepsilon_{ik} \mid y_{ik} > 0$$

where X_i denotes a vector of household characteristics and ε_{ik} the error term. This is implemented as a two-step estimation procedure with the first step utilizing a logit regression to estimate whether $y_{ik} = 0$ and the second step using an OLS regression. We use the framework of multiple imputations to obtain several point estimates by drawing from the error distribution to ensure accurate tails of the consumption distribution.

The household characteristics X_i are selected based on a step-forward algorithm minimizing the AIC by regressing household characteristics on the observed core and assigned non-core consumption including a fixed effect for the assigned module:

---

[6] Even though the consumption aggregate consists of complements and substitutes, a random allocation of items into optional modules will tend to make the correlation between modules positive except in the unlikely case of two modules sharing a large number of complements with one module capturing, for example, all of the items typically consumed by the poor. Thus, the optional modules can be assumed to be positively correlated.

$$\log(y_{i0} + y_{ik_i^*}) = \beta X_i + D_{ik} + \varepsilon_{ik} \mid y_{i0} + y_{ik_i^*} > 0$$

where D_{ik} represents *k* dummy variables with the k^{th} variable equal to 1 if household *i* is assigned to module *k* and equal to 0 otherwise.

Performance Assessment

We assess the performance of the estimation based on the bias and the coefficient of variation (CV). The bias is defined as the expected value of the absolute percentage difference of $\widehat{FGT}_{\alpha,z}$ estimated using the rapid approach and $FGT_{\alpha,z}$ estimated based on full consumption data. Using the additional index $1 \leq s \leq 20$ for the simulation, we obtain

$$E_z E_s \left| \widehat{FGT}_{\alpha,z,s} - FGT_{\alpha,z,s} \right|$$

Each simulation uses random allocations of items to optional modules and random assignments of households to optional modules.[7] We average the bias over all possible poverty lines *z* so that 1 percent, 2 percent, et cetera, and 99 percent of the population are defined as poor based on the full consumption distribution. The integration over all possible poverty lines makes the resulting performance measures independent of the poverty line, while the absolute difference in the definition of the bias avoids cancelling out errors.

Accordingly, the coefficient of variation is defined as the average ratio of the standard deviation and the mean of the FGT measure over all possible poverty lines:

$$E_z \frac{\sqrt{E_s \left(\widehat{FGT}_{\alpha,z,s} - FGT_{\alpha,z,s} \right)^2}}{E_s FGT_{\alpha,z,s}}.$$

Data

We applied this method to recent household consumption data from Kenya. Kenya's source for official poverty estimates is the Kenyan Integrated Household Budget Survey (KIHBS). The two last rounds were implemented in 2005/6 and 2015/16. The 2005/6 round used a representative sample of households in Kenya stratified by 7 provinces and 69 districts split into urban and rural. The sample size of the cleaned dataset includes 12,695 households in 1,338 clusters. The 2015/16 round used a representative sample of households in Kenya stratified by 47 counties split into urban and rural. The sample size of the cleaned dataset includes 21,585 households in 2,387 clusters. For both surveys, data collection was carried out over a period of 12 months.

The 2015/16 round was accompanied by a CAPI pilot implementing the rapid approach. The pilot used the same sampling frame as the 2015/16 round and interviewed up to an additional 6 households in the same clusters, resulting in a sample size of 12,670 households (as not all clusters and all intended households were interviewed due to non-response). The questionnaire was derived from the KIHBS 2015/16 questionnaire, but was considerably simplified across all modules. Specifically, the consumption module

---

[7] The constraints in these allocations are to ensure that items are uniformly distributed among optional modules and that each optional module is assigned equally often to households within each cluster.

was administered according to the rapid approach. We are using the dataset that was constructed with 5 food and 5 non-food items in the core (selected based on the highest consumption share in KIHBS 2005/6), with the remaining 128 food and 76 non-food items partitioned into 3 optional modules. Thus, the expected time savings were about 30 percent.

In the remainder of the paper, we use KIHBS 2015/16 with the full consumption module as a benchmark. The previous round of 2005/6 is used to define the core items, and for two of the alternative approaches it is used to adjust the poverty line and to build the consumption model for the cross-survey imputation. The 2015/16 pilot is used as a real-world example for the implementation of the rapid approach. We harmonize the datasets to ensure comparability across the different surveys. The harmonized datasets include 133 food and 81 non-food items. The consumption aggregate is exclusively based on these 214 items. In addition, we use harmonized location and household characteristics for the various models, including a binary and a categorical location variable as well as 6 additional binary, 9 additional categorical and 9 continuous variables for household characteristics (Table 4 in the Annex).

Consumption shares differ markedly between the 2005/6 and 2015/16 surveys, which is unsurprising given the 10-year gap (Kenya National Bureau of Statistics 2018). For example, the top ten food items in 2005/6 capture 52 percent of the food consumption share, but only 39 percent of the consumption share in 2015/16. Non-food consumption changed to a lesser degree. The top 10 items in 2005/6 represent 64 percent of non-food consumption shrinking to 59 percent in 2015/16. These differences are expected to impact the performance of those approaches that strongly rely on data from previous surveys, e.g. adjusting poverty lines and cross-survey imputations.

Results

In this section, we assess the performance of the rapid approach vis-à-vis alternative approaches, using the KIHBS 2015/16 full consumption data as a benchmark. First, we investigate the empirical tradeoff between the number of core items and the number of optional modules with respect to the performance of poverty indicators. Second, we compare the rapid approach with a traditional reduced consumption aggregate, without any adjustment of the poverty line. Third, we again use a reduced consumption aggregate, but adjust the poverty line based on previously observed consumption from 2005/6. Fourth, we compare the application of the rapid approach in the pilot in Kenya with a cross-survey imputation approach.

Trade-off with Number of Core Items and Optional Modules

As discussed in the methodology section, the rapid approach creates a tradeoff between the number of core items and the number of optional modules, as a larger number of core items (smaller number of optional modules) will improve the performance of the estimation. A larger number of core items will capture a larger fraction of the variance in the core module, minimizing the estimation error for the variance. Similarly, a smaller number of optional modules reduces the estimation error of the covariance between modules. However, the time savings by the rapid approach are larger for a smaller number of items in the core module and a larger number of optional modules as fewer items are asked about for each household.

Table 1: Trade-off for FGT measures by number of optional modules and number of core items measured by bias and cv.

	Opt.	Bias						CV					
		Core Items						Core Items					
	Mod.	0	1	3	5	10	20	0	1	3	5	10	20
fgt0	2	0.021	0.019	0.015	0.012	0.009	0.006	0.081	0.075	0.051	0.043	0.029	0.017
	4	0.033	0.032	0.024	0.019	0.014	0.007	0.125	0.122	0.085	0.07	0.051	0.026
	6	0.039	0.036	0.027	0.021	0.015	0.01	0.139	0.14	0.099	0.081	0.06	0.037
	8	0.042	0.042	0.031	0.026	0.021	0.009	0.151	0.155	0.105	0.096	0.08	0.035
fgt1	2	0.015	0.014	0.01	0.008	0.005	0.003	0.128	0.12	0.081	0.071	0.051	0.028
	4	0.024	0.023	0.016	0.013	0.009	0.005	0.19	0.189	0.129	0.114	0.087	0.045
	6	0.027	0.026	0.019	0.015	0.011	0.007	0.207	0.214	0.15	0.131	0.103	0.062
	8	0.029	0.029	0.02	0.018	0.015	0.006	0.224	0.232	0.156	0.151	0.132	0.063
fgt2	2	0.011	0.01	0.007	0.006	0.004	0.002	0.166	0.157	0.108	0.098	0.075	0.04
	4	0.017	0.017	0.012	0.01	0.007	0.004	0.241	0.241	0.166	0.154	0.123	0.064
	6	0.02	0.019	0.014	0.011	0.008	0.005	0.259	0.272	0.192	0.174	0.144	0.085
	8	0.021	0.022	0.015	0.013	0.011	0.005	0.279	0.29	0.198	0.198	0.179	0.09
gini	2	0.005	0.001	0.001	0	0.003	0.002	0.014	0.004	0.004	0.004	0.009	0.005
	4	0.004	0.001	0.003	0.001	0	0.003	0.013	0.009	0.01	0.004	0.004	0.01
	6	0.009	0.005	0.005	0.008	0.006	0.008	0.025	0.014	0.014	0.021	0.015	0.021
	8	0.001	0.008	0.011	0.003	0.006	0.007	0.011	0.022	0.03	0.01	0.015	0.018

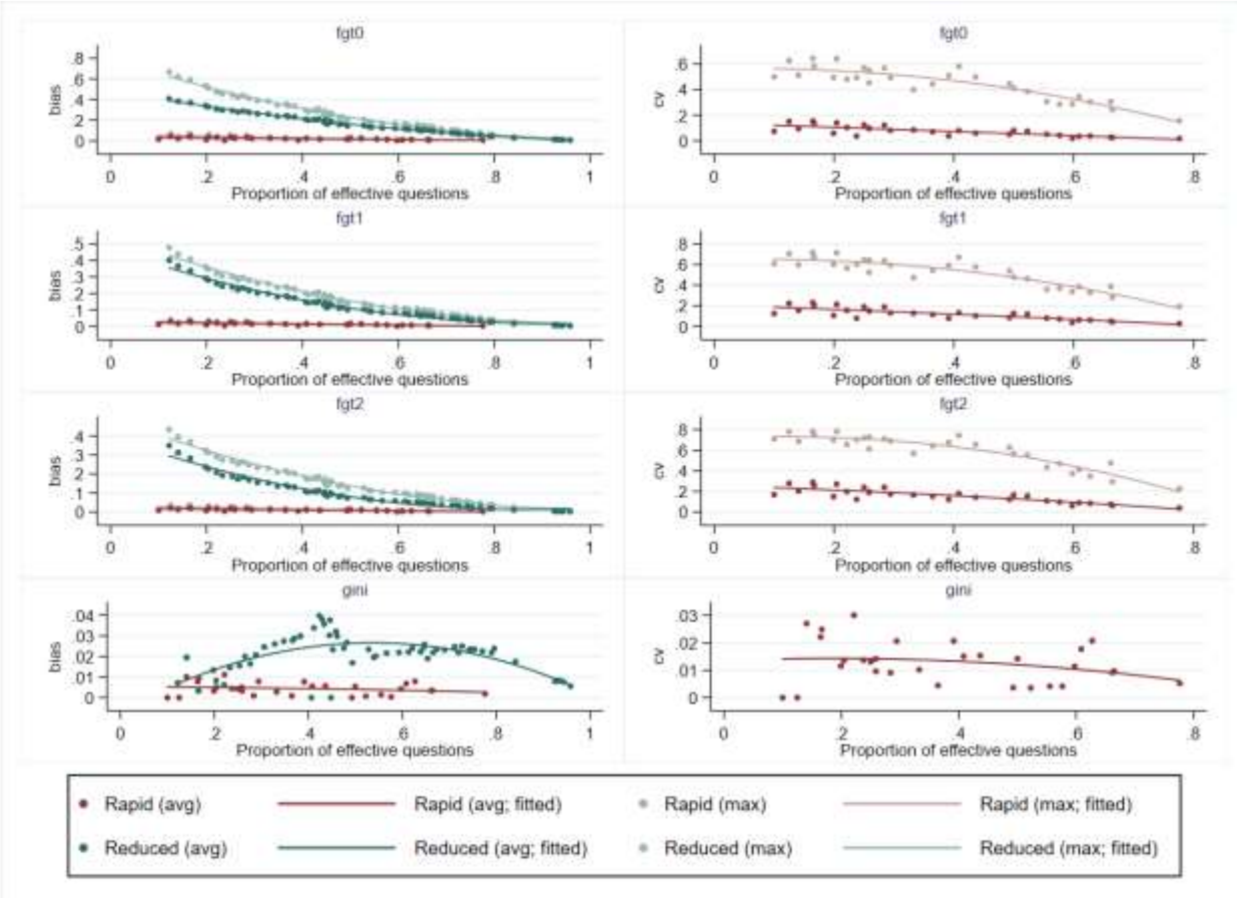
Based on simulations using KIHBS 2015/16, we estimate the bias and coefficient of variation to estimate FGT0, FGT1 and FGT2 as well as Gini using as reference the full consumption aggregate from the survey.[8] As expected, all performance measures deteriorate for a smaller number of core items, as well as for a larger number of optional modules (Table 1). Using the minimum number of optional modules (2), we obtain an average bias of 0.021 for FGT0 using no core items, but a considerably smaller bias of only 0.006 (a more than 70 percent reduction) if using 20 core items. An increase of the number of optional modules from 2 to 8 almost doubles the bias for FGT0 from 0.021 to 0.042 using 0 core items. We observe similar trade-offs for FGT1 and FGT2. The trade-offs for the Gini coefficient are less clear as independent of the number of core items and modules the bias is consistently extremely low almost always below 0.01.

Rapid vs. Reduced Estimation

Traditionally, time savings in administering consumption modules are achieved by reducing the number of consumption items included in the questionnaire. Here we compare the rapid approach with a reduced approach based on the effective time savings achieved. Assuming that only consumed items require substantial interview time, we estimate the average number of items that were administered and consumed by households relative to the total number of items in the full consumption module. The smaller the number of items consumed and administered in the questionnaire, the larger the time savings. This measure takes into account that effective time savings are smaller if fewer but often-consumed items are administered to a household, compared to a larger number of items, which are rarely consumed. The results show simulations with varying number of core items and optional modules for the rapid approach, and varying number of items included in the reduced module.

[8] If not noted otherwise, each core and optional module configuration is run 20 times, each using 50 multiple imputations. Note that the same survey data is used to define the core module items (based on highest consumption shares).

Figure 1: Performance of rapid vs. reduced poverty estimation, by effective questions asked



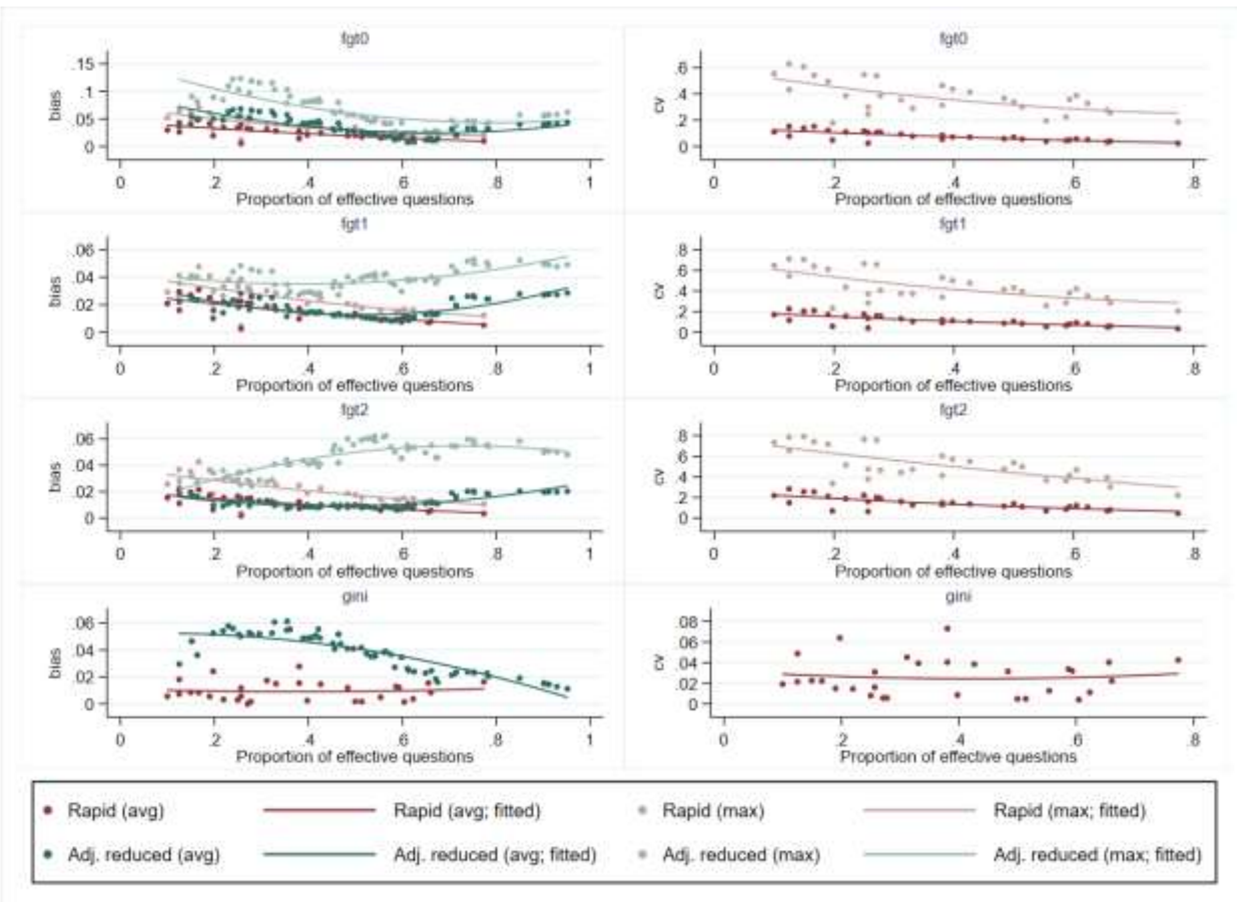
Based on KIHBS 2015/16 using the full consumption aggregate as the reference, we compare the average and maximum bias across all poverty lines for FGT0, FGT1 and FGT2 as well as Gini (Figure 1). The results clearly show the superiority of the rapid approach for any time saving larger than 10 percent. Generally, the bias and coefficient of variation are increasing for larger time savings, except for the Gini which is generally low with a bias of usually less than 0.01 for the rapid approach while the reduced approach over-estimates the Gini by up to 0.04. The maximum bias for FGT0 remains below 7 percent in 95 percent of the simulations. The average bias generally remains below 5.3 percent. The average coefficient of variation slightly increases for larger time savings, but generally remains below 20 percent. Note that the coefficient of variation is only meaningful for the rapid approach, as the reduced approach is deterministic across simulations.

Time savings of 50 percent can be achieved with the rapid approach by accepting an average bias of 1.8 percent and a maximum bias of 3 percent for FGT0 for 0 items in the core and 2 optional modules. For similar time savings, the reduced approach would need to consist of the 20 percent of items with largest consumption, which will be consumed by most households, but resulting in an average and maximum bias of 15 percent and 21.9 percent, respectively. A 75 percent time saving is possible with 0 core items and 4 optional modules. It comes at the cost of an average and maximum bias of 3.3 percent and 4.6 percent respectively.

Rapid vs. Adjusted Reduced Estimation

The reduced approach from the previous section can be improved by adjusting the poverty line based on a previous survey (Olson-Lanjouw and Lanjouw 2001). To simulate this case, we use the KIHBS 2005/6 survey to re-estimate the poverty line for the reduced approach and use the survey for the definition of core items for the rapid approach (optional module items are randomly assigned). The re-estimated poverty line for the reduced approach and the core module assignment is then applied to the KIHBS 2015/16 survey (Figure 2). As before, we only show the coefficient of variation for the rapid approach, as the adjusted poverty line approach is deterministic for each simulation.

Figure 2: Performance of rapid vs. Olson-Lanjouw and Lanjouw poverty estimation, by effective questions asked



We observe a very similar performance for the rapid approach as the performance does not change for any number of optional modules with zero core items. However, the best performance is now achieved with some core items, as it helps to isolate more of the variance from the estimation. For example, a time saving of 75 percent can now be achieved with a core module of 5 items and 10 optional modules resulting in an average and maximum bias of only 0.5 percent and 1.2 percent respectively for FGT0. This is a considerable reduction in bias as compared to using 0 core items and 4 optional modules (reported in the previous subsection), although in both cases the time savings are the same. Thus, it is useful to include a few core items with highest consumption shares, even if they are selected from a rather outdated survey as in Kenya with a 10-year gap.

The approach of using an adjusted poverty line performs significantly better than the simple reduced approach presented in the previous section, but at the cost of the distributional shape captured in the Gini coefficient. For the FGT measures, it is, thus, generally advisable to adjust the poverty line if a reduced approach must be used, even if the adjustment is based on outdated consumption shares from an old survey. For FGT0, the average bias of the adjusted reduced approach is usually around or above the maximum bias of the rapid approach. Furthermore, the maximum bias is considerably larger than the average bias for the adjusted, reduced approach compared to the rapid approach. For FGT1, the average bias of the adjusted, reduced approach becomes more comparable with the rapid approach, but the maximum bias is still significantly larger than the maximum bias of the rapid approach. For FGT2, the performance of the adjusted, reduced approach becomes more difficult to interpret. The Gini coefficient is not well conserved approaching a bias of 0.05 for highest time savings.

As the results show, the adjusted, reduced approach has larger variation across poverty lines (as shown by the larger difference between average and maximum bias) as well as across the different number of items (implying different time savings). The definition of the estimator explains this. The estimator depends strongly on the items selected for the reduced approach, as the accuracy relies on the adjustment factor of the poverty line, which is the share captured by those items. The error can be decomposed into two components. The first component is the change of the distributional shape between the survey used for the adjustment and the application of the adjusted poverty line. The second component depends on the difference in the share captured by those items between the two surveys. Both errors become zero if the approach is implemented for the same population at the same time. In the usual case though, neither the population nor the time point is the same. In these cases, especially the second component of the error leads to large variation of the performance. Even though the consumption shares of the items can change, the changes might cancel out leading to a good performance of the approach. However, adding one more item to the consumption module can void the cancellation, leading to a worse performance. Without knowing the share of the items from total consumption (which is not measured), it is impossible to predict how many items should be selected for a good performance of the approach.

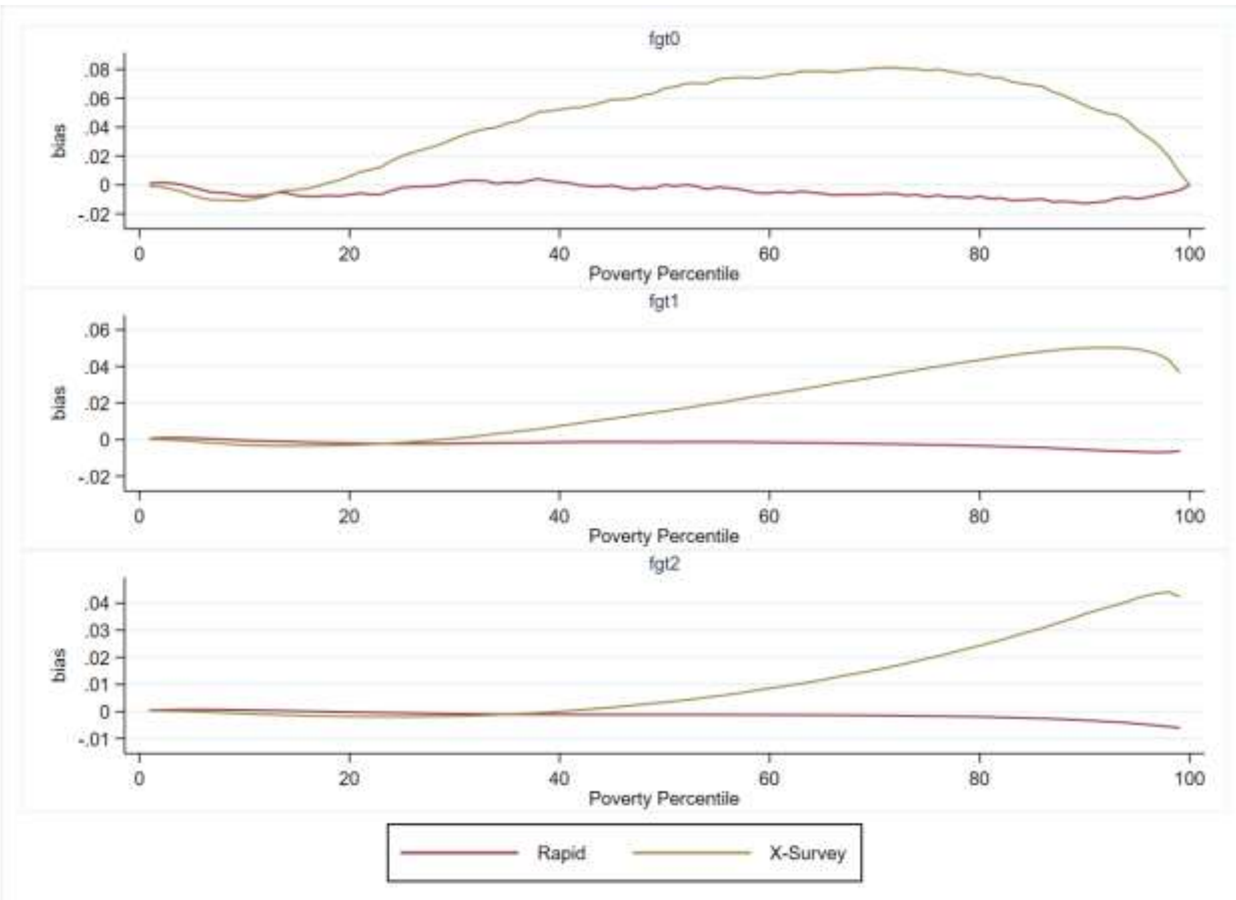
Application to Kenya: Rapid vs. Cross-Survey Estimation

In 2015/16, a CAPI pilot was implemented alongside KIHBS 2015/16 using the rapid approach. While the configuration is conservative with only 30 percent of time savings, the results show impressive performance for all FGT measures (Figure 3) in comparison with the full consumption as estimated for KIHBS 2015/16. Across all potential poverty lines, the rapid approach has a bias of below 1.3 percentage points for FGT0, 0.7 percentage points for FGT1 and 0.6 percentage points for FGT2. The Gini has a bias of only 0.012.

We compare the performance with a cross-survey imputation of consumption using a structural model built based on the KIHBS 2005/6 dataset and applied to KIHBS 2015/16, ignoring the collected consumption data in 2015/16. The performance of the structural model is then assessed against the KIHBS

2015/16 full consumption aggregate (Figure 3).[9] The cross-survey imputation cannot provide convincing results. FGT0 is under-estimated by up to 8.1 percentage points, FGT1 by up to 5.0 percentage points, and FGT2 by up to 4.4 percentage points. The Gini is off by 0.036. This is not surprising given the 10-year gap between the parameters of the structural model from 2005/6 and inference of poverty for 2015/16. In such long timeframes, not only do consumption patterns change but also structural drivers or correlates with poverty. To the further detriment of cross-survey imputations, it is in practice not possible to estimate the error, making it difficult to recommend its usage.

Figure 3: Bias of rapid vs. cross-survey (X-Survey) poverty estimation, by poverty percentile.



Conclusions

The rapid approach proposed in this paper can be used to achieve significant time savings, while only introducing a small bias into poverty estimates. The choice of the number of items in the core module and the number of optional modules allows for a precise calibration of time savings. The results show that it is helpful to utilize a previous survey from the same or a similar population to assign a few key items to

[9] The cross-survey imputation is based on the best model minimizing the AIC using a step-forward algorithm on the variables from KIHBS 2005/6. The results of the model selection are provided in Table 5 in the Annex. The imputation is performed in log-space with 50 multiple imputations.

the core module. In the best case, the selected items are still highly consumed and will improve the estimates of poverty. In the worst case, the selected items are no longer important in which case they will hardly affect the time savings compared to a design with zero core items and also equal its performance.

This paper also demonstrates the difficulty of achieving convincing results using alternative methods. Simply removing items from the consumption aggregate to create time savings, but without adjusting the poverty line, can lead to considerable bias in the poverty estimates. Better results can be achieved by adjusting the poverty line based on a previous survey in order to accommodate the reduced number of items, but without conserving the shape of the consumption distribution measured by the Gini. Furthermore, the potential for large changes in consumption patterns, which cannot be determined under this approach, creates considerable uncertainty in the resulting estimates. Similarly, a survey based only on covariates and their structural relationship with poverty estimates from a previous survey introduces a large bias in the estimates, at least in the studied case with a 10-year gap between surveys. The proposed rapid approach outperforms all these approaches, but at the cost of increased complexity.

The rapid approach introduces additional complexity into both the questionnaire design and the estimation of poverty. The capacity of enumerators is often low in developing countries. While the rapid approach increases the complexity of the questionnaire, CAPI technology easily solves this problem. Survey software can automatically compile a single consumption module based on the core and optional modules for each household, without making the partition explicit to the enumerator or demanding the execution of complex skip patterns. Furthermore, advanced CAPI technology can be used to generate the questionnaire automatically based on the assignment of the household to an optional module. While enumerators should be made aware that different households will be asked for different items, administering a rapid questionnaire does not require any additional training of enumerators beyond the standard skills for consumption questionnaires.

Conversely, the analysis of data using the rapid approach requires high analytical capacity, something that is usually lacking in developing countries. While the general concept of the assignment of optional consumption modules to households can usually be explained to local partners, poverty analysis based on a bootstrapped sample of the consumption distribution can potentially be too demanding for local capacity. However, even standard poverty analysis often surpasses the limits of local capacity, especially in conflict or post-conflict settings. Therefore, capacity building tends to focus on data collection skills with the long-term perspective of creating data analysis capacity. In addition, the rapid approach might be the only possibility to create poverty estimates in certain areas. For example in the case of Somalia, the rapid approach limited overall questionnaire administering time to less than 60 minutes for more than 90 percent of households as required by security considerations for enumerators (Pape and Mistiaen 2018; Pape and Wollburg 2019).

The rapid approach administers different consumption modules to different households. In theory, this can create a response bias if households report differently on a consumed item depending on the type and number of items previously asked. Unfortunately, we cannot estimate such a response bias in the available data. However, implementation of the rapid approach with an enhanced design with different optional modules varying in their comprehensiveness of items can in general shed light on this bias.

Comparison between responses for the same item in a comprehensive and a non-comprehensive list can also indicate a lower bound for response bias. Assuming that the context of a comprehensive list is a better estimate, the response bias could be corrected for. However, it is expected that this type of response bias is very small in comparison to general measurement and estimation errors.

The main source of bias for the rapid approach is created by the assumption of zero co-variance between optional modules. Further research can help to estimate co-variance between modules within the survey and adjust the consumption estimates accordingly. Using a random assignment of items to optional modules, the co-variance between groups of items within an optional module can potentially be used to estimate the co-variance between optional modules. Or, administering optional modules that share items might also be helpful to estimate the co-variance between modules.

The rapid approach reduces administering time considerably. While this creates opportunities to include additional questionnaire modules on different topics (e.g. remittances or health), it also has the potential to reduce the non-response rate. The KIHBS 2015/16 survey suffered from high non-response rate specifically in wealthier areas. For example, the capital city Nairobi had a response rate of 77 percent compared to 92 percent in the rest of the country. The highest non-response rates were specifically observed in clusters in wealthier areas within Nairobi. A high correlation of non-response with welfare status can lead to biased poverty estimates. The considerably shorter pilot survey – which not only used the rapid approach for consumption but generally shortened the questionnaire across modules and was carried out using tablets rather than paper – did not show the same pattern of lower response rates in wealthier areas. The pilot response rate of 99 in Nairobi was considerably higher than the standard KIHBS and was the same as in the rest of the country.[10] Thus, the rapid approach can help to contribute to shorter questionnaires mitigating concerns around low response rates, especially if correlated with welfare status.

The rapid approach might also be particularly useful in the context of evaluating shock or project impacts on poverty. In these cases, reliance on structural models estimated between surveys is dangerous. Shocks are likely to distort structural relationships between household characteristics and poverty. For example, a light shock is often mitigated by the household reducing its consumption, rather than selling assets or moving into another dwelling. A structural model estimated before the shock will not be able to capture the reduced consumption, thereby underestimating the impact of the shock. Similarly, project impacts cannot be adequately estimated using a structural model from before the project. For example, the distribution of metal sheets as rooftop materials is unlikely to change consumption patterns, but a structural model might use the absence of metal roofs to help predict poverty. While administering a full consumption module is often not feasible, especially in the case of shocks or in fragile settings, the rapid approach can readily be applied without relying on the assumptions of a structural model that would likely be violated.

[10] The change in the response rate is unlikely to be explained by the transition from PAPI to CAPI (Banks and Laurie 2000; Schräpler, Schupp, and Wagner 2010)

References

Banks, Randy, and Heather Laurie. 2000. "From Papi to Capi: The Case of the British Household Panel Survey." *Social Science Computer Review* 18 (4): 397–406. https://doi.org/10.1177/089443930001800403.

Beegle, Kathleen, Luc Christiaensen, Andrew Dabalen, and Isis Gaddis. 2016. *Poverty in a Rising Africa*. Washington, DC: World Bank Publications.

Beegle, Kathleen, Joachim De Weerdt, Jed Friedman, and John Gibson. 2012. "Methods of Household Consumption Measurement through Surveys: Experimental Results from Tanzania." *Journal of Development Economics* 98 (1): 3–18.

Christiaensen, L., P. Lanjouw, J. Luoto, and D. Stifel. 2011. "Small Area Estimation-Based Prediction Methods to Track Poverty: Validation and Applications." *Journal of Economic Inequality* 10 (2): 267–97.

Deaton, Angus, and Salman Zaidi. 2002. "Guidelines for Constructing Consumption Aggregates for Welfare Analysis." *Living Standards Measurement Study*, 135.

Douidich, M., A. Ezzrari, R. van der Weide, and P. Verme. 2013. "Estimating Quarterly Poverty Rates Using Labor Force Surveys." *World Bank Policy Research Working Paper*, no. 6466.

Foster, James, Joel Greer, and Erik Thorbecke. 1984. "A Class of Decomposable Poverty Measures." *Econometrica: Journal of the Econometric Society*, 761–66.

Fujii, Tomoki, and Roy van der Weide. 2016. "Is Predicted Data a Viable Alternative to Real Data?" *World Bank Policy Research Working Paper*, no. 7841 (October). https://papers.ssrn.com/abstract=2848469.

Graham, J. W., S. M. Hofer, and D. P. MacKinnon. 1996. "Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures." *Multivariate Behavioral Research* 31 (2): 197–218. https://doi.org/10.1207/s15327906mbr3102_3.

Kenya National Bureau of Statistics. 2018. "Basic Report on Well-Being in Kenya."

Olson-Lanjouw, Jean, and Peter Lanjouw. 2001. "How to Compare Apples and Oranges: Poverty Measurement Based on Different Definitions of Consumption." *Review of Income and Wealth* 47 (1): 25–42.

Pape, U, and J Mistiaen. 2015. "Measuring Household Consumption and Poverty in 60 Minutes: The Mogadishu. Washington DC: World Bank." Proceedings of ABCA Conference 2015. Washington DC: World Bank.

Pape, Utz Johann, and Johan A. Mistiaen. 2018. "Household Expenditure and Poverty Measures in 60 Minutes : A New Approach with Results from Mogadishu." *World Bank Policy Research Working Paper*, Policy Research Working Paper Series, , no. 8430. https://ideas.repec.org/p/wbk/wbrwps/8430.html.

Pape, Utz Johann, and Philip Randolph Wollburg. 2019. "Estimation of Poverty in Somalia Using Innovative Methodologies." *World Bank Policy Research Working Paper* no 8735 (February): 1–62.

Ravallion, Martin. 1998. "Poverty Lines in Theory and Practice." LSM133. The World Bank. http://documents.worldbank.org/curated/en/916871468766156239/Poverty-lines-in-theory-and-practice.

Schräpler, Jörg-Peter, Jürgen Schupp, and Gert G. Wagner. 2010. "Changing from PAPI to CAPI: Introducing CAPI in a Longitudinal Study." *Journal of Official Statistics* 26 (2): 239–69.

Appendix A: Performance of Estimation Techniques

Consumption of non-assigned optional modules can be estimated by different techniques. In addition to the two-step approach presented in the main text simple summary statistics and simple regression models can be used.

Summary Statistics (average and median)

This class of techniques applies a summary statistic on the collected module-specific consumption and applies the result to the non-administered modules. For each module k, a summary statistic $F\left(\left\{y_{jk} \mid j\colon k_j^* = k\right\}\right)$ can be computed based on households *j* to which the module *k* was administered so that consumption for household *i* can be estimated as

$$\hat{y}_{ik} = F\left(\left\{y_{jk} \mid j\colon k_j^* = k\right\}\right).$$

Using this approach, each household is assigned the same consumption per non-administered module. The summary statistics *F* can be, for example, a simple average or the median. The median has the advantage of being more robust against outliers but cannot capture small module-specific consumption if more than half of the households have zero consumption for the module.

Regression (OLS and Tobit regression)

Module-wise estimation applies a regression model for each module and exploits the differences in observed household characteristics

$$y_{jk} = \beta_k X_j + \varepsilon_{jk} \mid j\colon k_j^* = k$$

so that the deliberately absent consumption can be estimated as

$$\hat{y}_{ik} = \hat{\beta}_k X_i$$

with $\hat{\beta}$ representing the estimated OLS coefficient. Given the impossibility of negative consumption, a Tobit regression with a lower bound of 0 is used instead of a standard OLS regression approach. For the OLS regression, negative imputed values are set to zero.

Multiple Imputation

Single imputations of the consumption aggregate under-estimates the variance of household consumption. Depending on the location of the poverty line relative to the consumption distribution, this can either consistently under- or over-estimate poverty. Thus, the regression can also be embedded in a multiple imputation framework taking into account the variation absorbed in the residual term estimated via bootstrapping so that the resulting estimate becomes

$$\hat{y}_{ik} = \hat{\beta}_k X_i + \hat{\varepsilon}_k$$

where $\hat{\varepsilon}_k$ are repeated draws from the modeled residual distribution.

Performance Comparison

The comparison of the different estimation techniques reveals that the two-step estimation works well with highest consistency across different number of core items and different number of optional modules outperforming also the simple regression approach (Table 2 and Table 3).

Table 2: Performance by number of core items and estimation technique, using 2 optional modules.

	core items	0		1		3		5		10		20	
		bias	cv	bias	cv	bias	cv	bias	cv	bias	cv	bias	cv
FGT0	avg	0.16	0.55	0.16	0.55	0.14	0.51	0.12	0.45	0.1	0.4	0.06	0.29
	med	0.1	0.38	0.1	0.37	0.08	0.32	0.06	0.28	0.05	0.21	0.02	0.11
	mi_2cel	0.02	0.08	0.02	0.08	0.01	0.05	0.01	0.04	0.01	0.03	0.01	0.02
	mi_reg	0.05	0.4	0.05	0.4	0.05	0.39	0.04	0.33	0.04	0.31	0.03	0.24
	reg	0.03	0.09	0.03	0.09	0.02	0.06	0.01	0.04	0.01	0.03	0.01	0.03
	tobit	0.02	0.1	0.02	0.08	0.02	0.07	0.02	0.06	0.01	0.03	0.01	0.02
FGT1	avg	0.1	0.72	0.1	0.72	0.09	0.68	0.08	0.63	0.07	0.57	0.05	0.45
	med	0.05	0.5	0.05	0.48	0.04	0.44	0.03	0.4	0.03	0.32	0.01	0.18
	mi_2cel	0.02	0.13	0.01	0.12	0.01	0.08	0.01	0.07	0.01	0.05	<0.01	0.03
	mi_reg	0.06	2.57	0.06	2.53	0.06	2.38	0.05	1.79	0.04	1.58	0.03	1.09
	reg	0.02	0.12	0.02	0.1	0.01	0.06	0.01	0.04	0.01	0.05	<0.01	0.04
	tobit	0.01	0.15	0.01	0.1	0.01	0.09	0.01	0.08	<0.01	0.03	<0.01	0.01
FGT2	avg	0.08	0.81	0.08	0.81	0.07	0.78	0.06	0.74	0.05	0.68	0.04	0.56
	med	0.04	0.61	0.04	0.6	0.03	0.55	0.03	0.52	0.02	0.43	0.01	0.26
	mi_2cel	0.01	0.17	0.01	0.16	0.01	0.11	0.01	0.1	<0.01	0.08	<0.01	0.04
	mi_reg	0.12	13.98	0.12	13.52	0.11	12.16	0.08	7.92	0.07	6.6	0.05	3.94
	reg	0.01	0.14	0.01	0.1	0.01	0.07	0.01	0.05	<0.01	0.07	<0.01	0.06
	tobit	0.01	0.24	0.01	0.14	0.01	0.14	<0.01	0.12	<0.01	0.04	<0.01	0.01
Gini	avg	0.19	0.5	0.19	0.5	0.17	0.43	0.14	0.36	0.12	0.3	0.08	0.2
	med	0.15	0.39	0.15	0.39	0.13	0.32	0.1	0.27	0.08	0.21	0.05	0.12
	mi_2cel	0.01	0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	<0.01	0.01
	mi_reg	0.02	0.04	0.02	0.04	0.01	0.02	<0.01	0.01	<0.01	<0.01	0.01	0.02
	reg	0.03	0.08	0.03	0.08	0.02	0.06	0.02	0.04	0.01	0.03	0.01	0.02
	tobit	0.01	0.03	0.01	0.03	0.01	0.02	<0.01	0.01	<0.01	0.01	<0.01	<0.01

Table 3: Performance by number of optional modules and estimation technique, using 0 core items.

	opt. modules	2		4		6		8	
		bias	cv	bias	cv	bias	cv	bias	cv
FGT0	avg	0.16	0.55	0.23	0.67	0.25	0.7	0.26	0.71
	med	0.1	0.38	0.15	0.49	0.18	0.54	0.2	0.58
	mi_2cel	0.02	0.08	0.03	0.12	0.04	0.14	0.04	0.15
	mi_reg	0.05	0.4	0.07	0.48	0.07	0.48	0.07	0.5
	reg	0.03	0.09	0.04	0.11	0.05	0.16	0.05	0.17
	tobit	0.02	0.1	0.04	0.13	0.06	0.16	0.09	0.26
FGT1	avg	0.1	0.72	0.13	0.8	0.14	0.82	0.14	0.82
	med	0.05	0.5	0.06	0.52	0.06	0.52	0.07	0.53
	mi_2cel	0.02	0.13	0.02	0.19	0.03	0.21	0.03	0.22
	mi_reg	0.06	2.57	0.08	3.53	0.09	3.6	0.09	3.78
	reg	0.02	0.12	0.02	0.14	0.03	0.19	0.03	0.21
	tobit	0.01	0.15	0.02	0.2	0.03	0.18	0.05	0.32
FGT2	avg	0.08	0.81	0.09	0.88	0.09	0.89	0.1	0.89
	med	0.04	0.61	0.04	0.62	0.04	0.59	0.04	0.57
	mi_2cel	0.01	0.17	0.02	0.24	0.02	0.26	0.02	0.28
	mi_reg	0.12	13.98	0.18	22.71	0.19	23.59	0.2	25.36
	reg	0.01	0.14	0.01	0.16	0.02	0.21	0.02	0.23
	tobit	0.01	0.24	0.01	0.32	0.02	0.2	0.04	0.35
Gini	avg	0.19	0.5	0.28	0.73	0.31	0.81	0.33	0.85
	med	0.15	0.39	0.24	0.61	0.27	0.69	0.29	0.74
	mi_2cel	0.01	0.01	<0.01	0.01	0.01	0.02	<0.01	0.01
	mi_reg	0.02	0.04	0.02	0.06	0.03	0.07	0.03	0.07
	reg	0.03	0.08	0.04	0.1	0.05	0.13	0.06	0.15
	tobit	0.01	0.03	0.01	0.03	0.03	0.07	0.04	0.11

Appendix B: Additional Tables and Figures

Table 4: Harmonized household variables

Category	Variable	Type
Location	strata	categorical
	urban	binary
Household Characteristics	owns house	binary
	wall type	categorical
	roof type	categorical
	floor type	categorical
	improved drinking water source	binary
	improved sanitation facility	binary
	access to electricity	binary
	asset index from PCA	continuous
	quartiles of asset index from PCA	categorical
	number of rooms in household	continuous
	quartiles of number of rooms	categorical
	number of persons in household	continuous
	number of children in household	continuous
	proportion of children in household	continuous
	number of adults in household	continuous
	proportion of adults in household	continuous
	number of seniors in household	continuous
	proportion of seniors in household	continuous
	dependency ratio by intervals	categorical
	at least one member is literate 15+	binary
	male household head	binary
	household head age group	categorical
	household head education level	categorical
	household head employment type	categorical

Table 5: Model selection for rapid approach and cross-survey estimation.

	rapid		cross-survey	
dataset	KIHBS 2015/16 pilot		KIHBS 2005/6	
urban	0.0382	(1.71)	0.148***	(7.17)
owns house			-0.0453*	(-2.15)
wall type (category 2)			0.0454*	(2.24)
wall type (category 3)	0.0419*	(2.16)	0.112***	(5.77)
wall type (category 4)	-0.117***	(-5.18)		
wall type (category 5)	-0.0506*	(-2.46)	0.0730**	(2.74)
roof type (category 2)	-0.143***	(-4.27)	-0.106***	(-5.18)
roof type (category 3)	0.0985**	(3.08)	-0.0610*	(-2.54)
floor (category 2)	-0.157***	(-7.70)	-0.166***	(-9.33)
floor (category 3)	0.0673**	(2.98)	-0.111**	(-2.75)
improved drinking water source	0.0398**	(2.64)	0.0532***	(3.88)
improved sanitation facility	-0.0541**	(-3.02)	0.0671***	(5.17)
access to electricity	0.0383	(1.63)	0.0991***	(4.25)
asset index from PCA	0.0442**	(3.14)	0.110***	(19.98)
quartiles of asset index from PCA (2nd quartile)	0.0665**	(2.89)	0.0293	(1.46)
quartiles of asset index from PCA (3rd quartile)	0.118***	(3.69)	0.0349*	(2.37)
quartiles of asset index from PCA (4th quartile)	0.105	(1.95)		
number of rooms in household	0.0380***	(4.05)	0.0382***	(8.62)
quartiles of number of rooms (2nd quartile)	0.0524*	(2.22)	0.0263*	(1.97)
quartiles of number of rooms (3rd quartile)	0.0534	(1.76)		
quartiles of number of rooms (4th quartile)	0.0126	(0.25)		
number of persons in household	-0.00942	(-0.20)	0.194	(1.34)
number of children in household	-0.0562	(-1.17)	-0.226	(-1.56)
proportion of children in household	0.475*	(2.13)	-1.024***	(-16.40)
number of adults in household	-0.148**	(-3.06)	-0.352*	(-2.43)
proportion of adults in household	1.148***	(5.18)		
number of seniors in household	-0.179***	(-3.30)	-0.379**	(-2.60)
proportion of seniors in household	1.162***	(5.04)		
dependency ratio by intervals (2nd interval)			0.0539**	(3.07)
dependency ratio by intervals (3rd interval)			0.0237	(1.37)
at least one member is literate 15+			0.0528*	(2.44)
male household head	-0.0384*	(-2.30)		
household head age group (category 2)			-0.0413*	(-2.08)
household head age group (category 3)			-0.0322	(-1.40)
household head age group (category 4)			-0.0770**	(-2.68)
household head education level (category 2)	0.117***	(4.52)	0.0628**	(3.15)
household head education level (category 3)	0.134***	(4.54)	0.132***	(5.52)
household head education level (category 4)	0.186***	(5.24)	0.351***	(7.27)
household head employment type (category 2)	0.0632***	(3.30)		
household head employment type (category 3)	-0.0418	(-1.58)	-0.0778***	(-3.89)
household head employment type (category 4)	-0.135	(-1.34)	0.0227	(1.28)
assigned 2nd module	0.0224	(1.26)		
assigned 3rd module	-0.140***	(-7.88)		
constant	-0.587**	(-2.62)	1.352***	(34.65)
N	12658		12695	
R-sq	0.373		0.511	
adj. R-sq	0.371		0.509	
AIC	21502.0		20232.4	