Correcting Sampling and Nonresponse Bias in Phone Survey Poverty Estimation Using Reweighting and Poverty Projection Models

Kexin Zhang
Shinya Takamatsu
Nobuo Yoshida

Abstract

To monitor the evolution of household living conditions during the COVID-19 pandemic, the World Bank conducted COVID-19 High-Frequency Phone Surveys in around 80 countries. Phone surveys are cheap and easy to implement, but they have some major limitations, such as the absence of poverty data, sampling bias due to incomplete telephone coverage in many developing countries, and frequent nonresponses to phone interviews. To overcome these limitations, the World Bank conducted pilots in 20 countries where the Survey of Wellbeing via Instant and Frequent Tracking, a rapid poverty monitoring tool, was adopted to estimate poverty rates based on 10 to 15 simple questions collected via phone interviews, and where sampling weights were adjusted to correct the sampling and nonresponse bias. This paper examines whether reweighting procedures and the Survey of Wellbeing via Instant and Frequent Tracking methodology can eliminate the bias in poverty estimation based on the COVID-19 High-Frequency Phone Surveys. Experiments using artificial phone survey samples show that (i) reweighting procedures cannot fully eliminate bias in poverty estimates, as previous research has demonstrated, but (ii) when combined with Survey of Wellbeing via Instant and Frequent Tracking poverty projections, they effectively eliminate bias in poverty estimates and other statistics.

Correcting Sampling and Nonresponse Bias in Phone Survey Poverty Estimation Using Reweighting and Poverty Projection Models [*]

Kexin Zhang, Shinya Takamatsu and Nobuo Yoshida [a]

_____

a) Kexin Zhang, School of Agricultural Economics and Rural Development, Renmin University of China, zhangkexin629@ruc.edu.cn
Shinya Takamatsu, Poverty and Equity Global Practice, World Bank, stakamatsu@worldbank.org
Nobuo Yoshida, Poverty and Equity Global Practice, World Bank, nyoshida@worldbank.org

I.   Introduction and background

Phone interviewing opens new possibilities for empirical research in social sciences. It greatly reduces the costs of conducting surveys because all communication occurs without the face-to-face interaction customary in traditional surveys. With the outbreak of the COVID-19 pandemic limiting face-to-face interviews, phone surveys are more prevalent among academic institutions, survey companies, and individual researchers for individual-level and household-level data collection. To track households' living conditions on a timely basis during a pandemic, the World Bank launched the COVID-19 High-Frequency Phone Surveys (HFPS), which have been carried out in around 80 countries since March 2020. These surveys allow policy makers to monitor a wide variety of socioeconomic indicators in a timely and frequent manner.

Phone surveys have shortcomings. First, it is difficult to monitor poverty using the COVID-19 HFPS. The COVID-19 HFPS does not collect consumption or income data, which are necessary for measuring poverty and inequality under the traditional poverty monitoring approach. It is time-consuming, costly, and complex to collect such data. An interviewer must ask households a number of complex questions regarding their recent consumption, expenditures, and detailed income components. Such an interview requires at least 30 minutes and up to two hours, even by well-trained and qualified interviewers. It is also challenging to administer the interview in developing countries where telephone connections are not always stable enough to complete such a long interview.

A solution to Challenge 1: SWIFT as a rapid poverty monitoring tool

In March 2020, the World Bank launched a pilot for the use of a rapid poverty monitoring tool, Survey of Wellbeing via Instant, Frequent Tracking (SWIFT), to estimate monetary poverty using the COVID-19 HFPS. SWIFT adds 10 to 15 simple questions to the COVID-19 HFPS questionnaire, and these additional questions take 3 to 5 minutes to ask. Based on households' responses to these questions, the SWIFT method can be adopted to impute poverty and inequality rates among the population of interest.

SWIFT was chosen among many poverty projection methodologies because it has been thoroughly tested and reviewed by experts inside and outside the World Bank. The results are encouraging – the difference between the true and SWIFT-based poverty projections is 1.2 percentage points on average and is within +/- 2 percentage points, except for one case (see more details in Appendix 5 or Yoshida et al., 2022a). The pilot was conducted in 20 countries around the world, and the results of seven countries are summarized in Yoshida et al. (2022b).

Second, the sample of the COVID-19 HFPS is unlikely to be nationally representative. In many developing countries, the coverage of mobile phones or land lines is limited. Poor households that do not own a phone are excluded from phone interviews. In addition, phone interviews usually face a much higher rate of nonresponse than in-person interviews. Such nonresponses are usually concentrated among urban residents who are too busy to participate in the interviews. Hence, the sample collected in a phone survey is often far from nationally representative and can result in bias in the projections of poverty.

A solution to Challenge 2: Reweighting – sampling weight adjustments to address sampling biases

Many approaches for adjusting sampling weights, or "reweighting," have been developed to correct for various types of sampling bias. This paper classifies them into two major categories: (1) propensity score weighting (PSW) and (2) "non-PSW adjustments."

Propensity score weighting is a method for adjusting sampling weights based on propensity scores and was developed from the propensity score matching proposed by Rubin and Rosenbaum (1983 and 1984). In the context of causal inferences, propensity score matching makes the control and treatment groups comparable, minimizing bias in estimating treatment effects. Unless the samples of the control and treatment groups are selected randomly (which is usually not the case with observational data), baseline characteristics may exhibit

differences between the two groups. Propensity scores are often estimated to address this issue of non-comparability. A propensity score is the probability of selecting a sample household or individual into the treatment group from the population, conditional on covariates. If a set of assumptions is satisfied (see Rosenbaum and Rubin 1983 and 1984 for details), a comparison of the weighted average of the outcome indicators between the control and treatment groups provides an unbiased estimate for the treatment effect.

Taylor (2000) and Lee (2006) adopted the propensity score matching technique to adjust sampling weights and correct for sampling bias in a web survey. First, they chose a reference survey representative of the population of interest (e.g., the entire population of a country, the urban population, or all refugees in a country). They combined the reference and web surveys and estimated propensity scores using this combined sample. They then divided the combined sample into quintiles based on the propensity scores and adjusted the sampling weights of the web survey to equate the sum of the weights of the web survey with those of the reference survey in each quintile.

Another type of propensity score weighting method is "inverse probability approach," which uses propensity scores to construct the odds of a sample being selected for a reference survey, and the odds are further used as a sampling weight of a web or phone survey (Morgan and Todd 2008, Schafer and Kang 2008, and Austin 2011). This paper classifies these types of reweighting techniques based on propensity scores as Propensity Score Weighting (PSW) methods. If the assumptions in Rubin and Rosembaum (1983 and 1984) are satisfied, then summary statistics applying PSW-based weights in a web or phone survey become representative of the population of interest.

A second category of reweighting techniques, which we refer to as non-PSW approaches, matches indicators between a phone or web survey and a reference survey at a more aggregate level than at the household/individual level. This group of approaches includes raking, post-stratification, and maxentropy. The non-PSW approaches select a set of indicators and adjust sampling weights so that weighted averages of the selected indicators are closely/exactly matched between a phone/web survey and a reference survey.

PSW and non-PSW adjustments are often conducted together. We consider two advantages of the combined applications of PSW and non-PSW approaches. First, for PSW to eliminate sampling bias, a set of assumptions, like strong ignorability, need to be satisfied, which cannot be easily tested, for which non-PSW serves as a complement for weighting purposes. Second, non-PSW approaches match the means of key indicators between a phone/web survey and a reference survey, but there is no guarantee that the distributions of these indicators are also matched. PSW approaches, by contrast, match the distribution of propensity scores. Since both approaches have their own strengths, conducting both PSW and non-PSW adjustments is reasonable.

This paper investigates whether reweighting can correct the bias of poverty projections produced by the SWIFT methodology. The performance of reweighting techniques differs by data and target indicators that were matched, and there is agreement in the literature that reweighting techniques reduce the biases in target statistics yet do not eliminate them (Lee (2006) and Dreze and Somanchi (2023)). Dreze and Somanchi (2023) created biased samples by dropping poorer households from a household survey and tested whether a non-PSW reweighting technique (maximum entropy reweighting, or maxentropy) can reduce biases in poverty rates and mean household expenditures. Although the biases in poverty rate estimates and means of household expenditures declined, substantial proportions remained.

However, existing literature lacks an assessment of how well reweighting techniques can reduce the biases of poverty projections produced by SWIFT or any other poverty projection method. Using phone or web surveys to estimate poverty necessitates the use of poverty projection methods. Dreze and Somanchi (2023) used actual consumption and income data and showed that a large bias in the poverty rate and mean household expenditure remains even after reweighting but did not assess if reweighting combined with poverty projection methods is

effective in reducing the bias. In fact, our study finds that the performance of reweighting in estimating poverty can be improved when combined with poverty imputation methods such as SWIFT.

Experiments

This paper examines whether, and if so, to what extent, a combination of reweighting techniques and the SWIFT poverty projection methodology can eliminate sampling biases in poverty estimates based on biased survey samples. To see this, we first use reference household surveys in Rwanda, St Lucia, and Uganda and construct subsamples by selecting households with at least a mobile or landline phone. These samples are subject to sampling bias by construction. Without reweighting and SWIFT poverty projections, the poverty rates in these subsamples of phone owners are lower than those in the full samples. We then examine whether reweighting and SWIFT poverty projections can correct for the abovementioned bias.

Phone and web survey data collections face sampling and nonresponse biases, but the abovementioned experiments only focus on sampling biases that arise from uneven phone ownership. To understand the ability of the SWIFT and reweighting techniques to adjust for nonresponse bias, this paper conducts an additional experiment using the sample of Ethiopia High-Frequency Phone Survey round 7 (HFPS7), which is a subsample of the 2018/19 Ethiopia Socioeconomic Survey round 4 (ESS4). Since this subsample of ESS4 includes only phone owners, it is subject to sampling bias. Also, since it includes only households of ESS4 that responded to the HFPS7, it is also subject to nonresponse bias. Using this subsample, we conduct the same analysis as above to assess whether reweighting techniques and SWIFT can correct the bias in poverty estimates arising from both sampling and nonresponse biases.

This paper is organized as follows. Section II introduces the SWIFT poverty projection methodology and reweighting techniques, and Section III displays results from four experimental studies (Saint Lucia, Rwanda, Uganda, and Ethiopia). Section IV concludes with an assessment of a combination of reweighting and the SWIFT-based poverty projection model in eliminating the bias in poverty estimation based on the four case studies.

II.   SWIFT and Reweighting Methodology

II.1. SWIFT poverty projection methodology

SWIFT is an application of Survey-to-Survey imputation techniques (S2S) to monitor poverty rapidly. SWIFT trains an imputation model in a training dataset by regressing household expenditures/incomes on poverty proxies. Household expenditures and poverty rates are then imputed in another dataset, called "output data," by plugging poverty proxies of the output data into the model. Figure 1 illustrates the process.

There are two key assumptions in the standard SWIFT methodology. First, the relationship between household income or expenditure and poverty correlates in the output data can be expressed in equation (1):

$$lny_{ho} = x_{ho}'\beta_o + u_{ho} \quad (1)$$

$$\text{where } u_{ho} \sim N(0, \sigma_o)$$

where $ln\, y_{ho}$ refers to a natural logarithm of the household income or expenditure of household *h* in the output data *o*. x_{ho} is a $(k \times 1)$ vector of poverty correlates of household *h* in the output data, *o*. β_o is a $(k \times 1)$ vector of coefficients of poverty correlates (x_{ho}). u_{ho} refers to a residual and is often assumed to follow a normal distribution of $N(0, \sigma_o)$.[1] The output data includes the poverty proxy data $\{x_{ho}\}_{h=1}^{H}$ but does not include

---

[1] This normal distribution and linearity can be relaxed. For the sake of exposition, normal distribution is assumed.

household expenditures $\{lny_{ho}\}_{h=1}^{H}$, which are to be imputed. For the sake of exposition, the relationship is assumed to be linear, but this condition can be relaxed (Yoshida et al., 2022a).

Figure 1 Illustration of how the SWIFT works

	training data		output data

Dataset	$\{lny_{ht}, x_{ht}\}_{h=1}^{H}$	----------------------	$\{x_{ho}\}_{h=1}^{H}$
Relationship	$lny_{ht} = x_{th}\beta_o + u_{ht}$	---------------------- Model stability holds	$lny_{ho} = x_{ho}\beta_o + u_{ho}$
Imputation	--		$\widetilde{lny}_{ho}^{r} = x_{ho}\tilde{\beta}^{r} + \tilde{u}_{ho}^{r}$
Imputed data	--		$\{\{\widetilde{lny}_{ho}^{r}, x_{ho}\}_{h=1}^{H}\}_{r=1}^{100}$
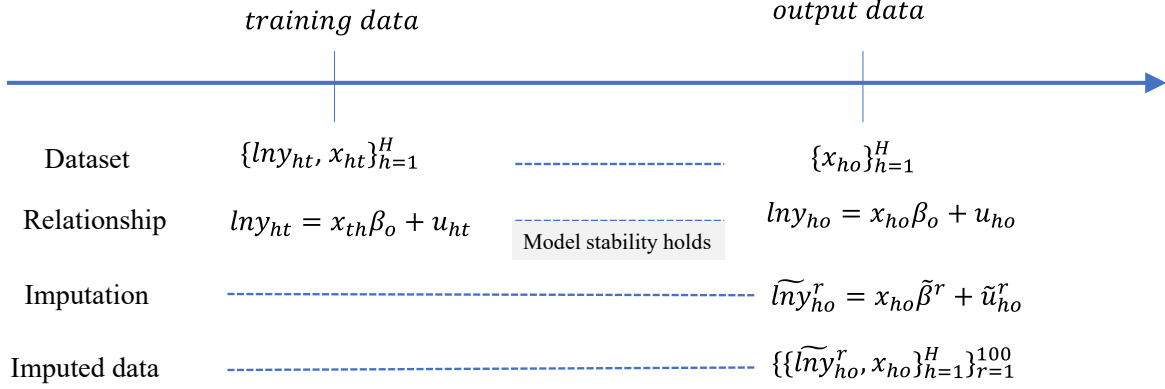
Note: Authors' illustration.

The second key assumption is that the relationship between household expenditures and poverty proxies in the training data also follows equation (1). This assumption is called "model stability" implying that the model does not change from the training and output data.

SWIFT estimates parameters in equation (1'), $(\hat{\beta}_o, \hat{\sigma}_o)$, and their distributions, using the training data set, draw them $(\tilde{\beta}^{r}, \tilde{u}_h^{r})$ randomly from their estimated distributions, and substitutes them into equation (1) to impute household expenditures for all households in the output data. SWIFT repeats this imputation (typically 20 – 100 times), resulting in 20 to 100 vectors of household expenditures $(\widetilde{lny}_{ho}^{r})$ in the output data (with each vector including the expenditure for all households). Poverty and inequality measures are estimated in each of the 20 – 100 vectors, and the averages are the point estimates of poverty and inequality. Standard errors can also be computed using the 20 to 100 estimates.

Yoshida et al. (2022a) evaluates how restrictive the above-mentioned assumptions are. First, they find that the use of a linear model is not mandatory. They compare the performance of linear models with that of non-parametric models and find that differences are minimal. Second, the model stability assumption can hold even after a large shock, like the COVID-19 pandemic, occurs. It is often argued that households' livelihoods and living conditions are likely to have been impacted by a large shock like the COVID-19 pandemic and the model trained by the pre-shock dataset no longer reflects the post-shock environment. As a result, poverty rates estimated by the pre-shock model can be biased. However, Yoshida et al. (2022a) find that if fast-changing poverty correlates are included in the model, the bias in poverty rates can be minimized. In the SWIFT-COVID-19 pilot, this approach – inclusion of fast-changing poverty correlates – is adopted (see Yoshida et al. 2022b).

II.2. Propensity Score Weighting (PSW)

A propensity score is the probability of a sample being selected into the "treatment group" from combined data of "treatment" and "control" groups. Propensity score matching was first developed by Rubin and Rosenbaum (1983 and 1984) and enabled researchers to estimate an unbiased treatment effect when the selection of either the control group or the treatment group is not randomized. Specifically, propensity score matching allows one to balance the distributions of covariates in the control and treatment data (Rubin and Rosenbaum, 1983; Schafer and Kang, 2008). In contrast, propensity score weighting (PSW) uses propensity scores to balance samples of a biased survey, like a phone or web survey, with those of a reference survey, which is known to be representative of a population of interest *ex-ante* (Taylor 2000 and Lee 2006). PSW has

been widely employed to correct bias in survey estimates arising from noncoverage problems (Duncan and Stasny 2001), late response (Czajka et al. 1992), and nonresponse (Smith et al. 2000; Little and Vartivarian 2003).

This paper examines two PSW approaches. The first approach follows the concept of stratification or blocking, where a combined dataset of a reference survey and a biased survey is split into multiple (usually five) groups based on the estimated propensity scores (Schafer and Kang 2008 and Austin 2011). Samples of both surveys that fall into the same group are considered comparable. Lee (2006) adjusts the weights of the phone/web survey in a way that the sum of weights in the phone/web survey sample in each group is equal to the counterpart of the reference survey. The second PSW approach uses inverse-propensity scores to adjust bias (for example, Morgan and Todd 2008, Schafer and Kang 2008, and Austin 2011). This approach multiplies the original sampling weight of a phone survey with the odds of a sample being selected into the reference survey from the combined data set. Both of the abovementioned PSW approaches aim to eliminate sampling bias to make summary statistics of the phone/web survey comparable to those of the reference survey and thereby make the adjusted phone/web survey representative of the population of interest. Below we present each PSW method in more detail.

PSW following Lee (2006)

We first revisit the propensity score weighting procedure proposed by Lee (2006). Suppose that there are two samples: (a) a phone survey sample (s^p) with n^p units with a base weight of d_j^p where $j = 1, ..., n^p$, and (b) a reference survey (usually a nationally representative household survey) sample (s^R) with n^R units with a base weight of d_j^R where $j = 1, ..., n^R$. First, the two samples are combined into one, $s = (s^p \cup s^R)$ with $n = n^p + n^R$ units. Propensity scores are calculated from the combined data. The propensity score of the ith unit is the likelihood of the unit participating in the phone survey ($g = 1$) as opposed to the reference survey ($g = 0$), where $i = 1, ..., n$. Propensity scores are defined as $e(x_i) = \Pr (i \in s^p | x_i, i = 1, ..., n)$ and are estimated in a logistic regression as in the following equation using x_{obs}, covariates observed commonly in the phone and the reference survey.

$$ ln \left[\frac{e(x)}{1 - e(x)} \right] = \alpha + \beta^T f(x) $$

where $f(x)$ is some function of the covariates. Critical assumptions underlying this practice are that: (a) for a given set of covariate values, a person/household must have some nonzero probability of being in the phone survey and that (b) probability must be estimable from the combined sample, s.

Based on the predicted propensity score, $\hat{e}(x_{obs})$, the distribution of the phone sample units is rearranged so that s^p (phone sample) resembles s^R (reference sample). Mechanically, this is first done by sorting s by $\hat{e}(x_{obs})$ and portioning s into c subclasses. A conventional choice is to use five subclasses based on quintile points (Cochran 1968). Ideally, all units in a given subclass will have approximately the same propensity score or the range of scores in each class will be fairly narrow. In the cth subclass, denoted as s_c, there are $n_c = n_c^p + n_c^R$ units, where n_c^p is the number of units from s^p, and n_c^R from s^R. The total number of units s remains the same:

$$ \sum_{c=1}^{C} (n_c^p + n_c^R) = \sum_{c=1}^{C} n_c = n $$

Second, the following adjustment factors are computed for each subclass c:

$$ f_c = \frac{\sum_{k \in s_c^R} d_k^R / \sum_{k \in (s^R)} d_k^R}{\sum_{j \in s_c^p} d_j^p / \sum_{j \in s^p} d_j^p} $$

where s_c^R and s_c^p are the sets of observations of the c th subclass in the reference and phone samples, respectively; s^R and s^p full samples of the reference and phone surveys; d_k^R represents the weights in the k-th group of the reference survey; d_j^p the weights in the j-th group of the phone survey.

The adjusted weight for unit j in class c of the phone survey sample becomes

$$d_j^{p.PSA} = f_c d_j^p$$

The estimator for the mean of a study variable, y, from s^p becomes

$$\bar{y}^{p.PSA} = \frac{\sum_c \sum_{j \in (s_c^p)} d_j^{p.PSA} y_j}{\sum_c \sum_{j \in (s_c^p)} d_j^{p.PSA}}$$

PSW using inverse-propensity scores (Schafer and Kang 2008 and Austin 2011)

In this section, we follow Schafer and Kang (2008) and Austin (2011) to explain the inverse-propensity-score approach in constructing weights. In the context of propensity score matching, we can make the average outcome of the control group comparable to that of the treatment group by multiplying the initial sampling weight of the control group by the odds of a sample being selected for the treatment group.

Suppose the original sampling weight of the control group is w_i^c, the propensity score for the sample i to be selected into the treatment group p_i , and the control group 1-p_i . Then, a new weight to make the sample i in the control group comparable to the treatment group counterpart is:

$$w_i^c * \frac{p_i}{1 - p_i}$$

Note that the second term is the odd for selecting this sample from the control group into the treatment group. The average of outcomes weighted by this new sampling weight in the control group is now comparable to that of the treatment group (see more details in Schafer and Kang 2008 or Austin 2011).

By the same token, this approach can be used to reweight the phone survey. Now, suppose the original sampling weight of the phone survey is w_i^p and the propensity score of sample i to be selected for the phone survey is p_i , and the reference survey $1 - p_i$. Then, a new weight to make sample i in the phone survey comparable to the counterpart in the reference survey is:

$$w_i^p * \frac{1 - p_i}{p_i}$$

The second term is the odd for sample i in the phone survey to be selected for the reference survey. The average outcomes weighted by this new sampling weight in the phone survey are now comparable to that of the reference survey (see formal proofs in Schafer and Kang 2008 and Austin 2011).

Selection of covariates for the estimation of propensity scores
Propensity score weighting is feasible when the probability predictors are available from both the reference and phone surveys.[2]

---

[2] For the units where some of the covariates are missing, propensity scores cannot be computed, which hinders the adjustment (Lee 2006).

The set of covariates typically includes similar demographic variables to those used in traditional post-stratification adjustment. Harris Interactive, a market research agency, includes demographic and non-demographic variables in the propensity models (Terhanian et al., 2000; Taylor et al., 2001). Rubin and Thomas (1996) suggested including all covariates, even if some are not statistically significant in predicting the outcome variables unless they are unrelated to the treatment outcomes or inappropriate for the model. By the same token, Li, Morgan, and Zaslavsky (2018) proposed that a rich propensity score model, rather than a parsimonious one, is desirable, especially in causal inference applications, because the ignorability assumptions are likely to be violated if the propensity score is only a simple logistic-linear function of the covariates, but both the outcome and assignment mechanism are functions of more complex interactions or nonlinear terms. On the other hand, Lee (2006) suggested that the importance of including non-demographic variables in PSW for web surveys is unclear due to two facts: (1) the inclusion of more variables automatically increases the predictive power of the model, but (2) non-demographic covariates can often be explained by demographic variables, to some degree.

While there has not been a rule of thumb on what variables should be included in the PSW procedure, in the simulation study by Drake (1993), the impact of specifying incorrect propensity score models, such as mistakenly adding a quadratic term or dropping a covariate, was not very serious. In fact, the misspecification of the propensity score model led to only a small bias compared to the misspecification of the response model that was used to simulate the response distribution (Drake 1993).

However, if the reference survey was collected a few or more years before the phone survey, it is recommended not to include covariates that are likely to change drastically over time. PSW balances the phone survey with the reference survey using propensity scores estimated with household-level covariates from both surveys. Some covariates, such as asset ownership or employment conditions, may have changed over the period. For example, if a country grows rapidly, asset ownership and employment conditions can change greatly in only one to two years. If these time-varying indicators were included in the propensity score regression, PSW would match these indicators as if they were constant over the interval between the reference and phone surveys. Such false matching could cause a significant bias in poverty projections. Therefore, we recommend using time-invariant covariates for the propensity score regression, particularly if the interval between the phone and reference surveys is long. Variables that are usually included in the PSW procedure are household size, dependency ratio, and household locality (urban/rural). If applicable, additional variables such as the highest educational attainment of the household head, age of the household head, sex of the household head, geographical characteristics of the household, and dwelling characteristics of the household can also be included in the PSW procedure as predictors of a household's probability of being reached by phone.

II.3. Non-PSW adjustments

PSW has good statistical properties in correcting for sampling biases by adjusting sampling weights to match the distribution of target indicators at the household level. Nevertheless, even with PSW-based weight adjustments, the means of time-invariant indicators, such as average household size, can still differ between a phone survey and a reference survey. PSW makes the two surveys comparable, but not perfectly so; some presumably time-invariant variables may still exhibit significant differences in means between the reweighted phone survey and the reference survey. Such remaining non-negligible differences can also cause bias in poverty projections.

Another limitation of PSW is that it cannot be executed if household unit records are unavailable in reference data. Although some statistics representative of a population interest are easily available from official sources, the source data may not be readily available. For example, regional population shares are updated frequently by the Bureau of Statistics of many countries, but the source data is not readily available on a timely basis. Under this circumstance, PSW cannot be used for reweighting.

These limitations motivate additional sampling adjustments, namely, the non-PSW adjustments discussed in this paper. This section discusses the three most commonly adopted non-PSW methodologies in the literature: maxentropy, raking, and post-stratification. These methodologies can be implemented in the absence of household unit records as long as the aggregate summary statistics are available. These approaches can also eliminate any remaining mismatches in means after PSW adjustments have been performed.

<u>Maximum Entropy Reweighting (or Maxentropy)</u>

Maxentropy (Golan, Judge, and Miller 1996) aligns the means of time-invariant variables between the weighted phone survey and the reference survey exactly. The idea of maximum extropy estimation was motivated by Jaynes (1957) to address the problem of finding the probability distribution $(p_1, p_2, ..., p_n)$ for a set of values $(x_1, x_2, ..., x_n)$, given only their expectation,

$$E\{f(x)\} = \sum_{i=1}^{n} p_i f((x_i)$$

Where $f((x_i)$ is the probability density function of x_i. To illustrate how this works, we borrow an example from Wittenberg (2010).[3] Consider a die known to have $E(x) = 4$, where x = (1, 2 ,3 ,4 ,5, 6), and we want to determine the associated probabilities. Obviously, $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$ cannot be a solution to this problem. Jaynes's solution was to tackle this from the point of view of Shannon's information theory. Jaynes wanted a criterion function $H(p_1, p_2, ..., p_n)$ that would characterize the uncertainty about the distribution. This is uniquely given by the entropy measure,

$$H(p_1, p_2, ..., p_n) = -K \sum_{i=1}^{n} p_i \ln(p_i)$$

where $p_i \ln(p_i)$ is defined to be zero if $p_i = 0$ for some positive constant K. The solution to Jaynes's problem was to pick the distribution $(p_1, p_2, ..., p_n)$ that maximizes the entropy, subject to the constraints,

$$E\{f(x)\} = \sum_{i} p_i f(x_i)$$

$$\sum_{i} p_i = 1$$

As Golan, Judge, and Miller (1996) showed, if our knowledge of $E\{f(x)\}$ is based on the outcome of N (very large) trials, then the distribution function $\boldsymbol{p} = (p_1, p_2, ..., p_n)$ that maximizes the entropy measure is the distribution that can give rise to the observed outcomes in the greatest number of ways, which is consistent with what we know from Wittenberg (2010).

More formally, the maximum entropy problem can be represented as:

$$\max_{p} H(\boldsymbol{p}) = -\sum_{i=1}^{n} p_i \ln(p_i)$$

$$such\ that\ y_j = \sum_{i=1}^{n} X_{ji} p_i, j = 1, ..., J \qquad (1)$$

---

[3] Details are available in Appendix 4.

$$\sum_{i=1}^{n} p_i = 1 \qquad (2)$$

Variables in the matching conditions (1) need to be time-invariant between the reference and phone surveys if the phone survey is carried out a few or more years after the reference survey. If the socioeconomic environment is changing rapidly and the summary statistics of some variables are likely to change after the reference survey was conducted, these variables should not be included in the matching conditions as if they were constant. Therefore, we recommend including only time-invariant or slowly changing variables in the matching conditions. For example, the average household size or the dependency ratio usually takes time to change. Also, housing conditions cannot be changed quickly, even if a large shock like the COVID-19 pandemic occurs.[4] On the other hand, employment conditions and economic sentiments can change very quickly and are therefore not recommended to be included as matching conditions for the maxentropy command. [5]

Raking

Raking, also referred to as "iterative proportional fitting," is an algorithm that consists of an outer cycle that checks convergence criteria and an inner cycle that iterates over the control variables. Once the phone survey with updated weights shows convergence along a certain control variable, the algorithm moves to the next control variable and adjusts the weights until convergence on that variable. The process continues until the weights achieve convergence along all specified variables (Kolenikov 2014).

Capacci et al. (2018) recommended raking after PSW to balance the means of key variables between a survey of interest and a reference survey. This subsection introduces raking as a way of balancing the means of key variables after PSW is conducted.

Like maxentropy, variables included in the raking procedure need to be time-invariant between the reference and phone surveys. Some examples include household size (discretized into 3 to 4 categories), dependent share (discretized into 4 to 5 categories), household locality (rural/urban), highest educational attainment of household head, gender of household head, age of household head, geographic variables, and time-invariant asset/dwelling characteristics (wall, roof, fuel, toilet conditions, etc.).[6]

For the detailed Stata codes for implementing raking, we recommend referring to Kolenikov (2014), which describes this procedure in more detail. The Stata codes are readily available as user-written commands.

---

[4] Some potential candidates for the maxentropy procedure include, but are not limited to, polynomials of household size, dependency ratio, household locality (urban/rural), the highest educational attainment of the household head, age of household head, sex of household head, geographical characteristics, dwelling characteristics (roof/wall/fuel type/toilet type), and occasionally assets that cannot be easily replaced (refrigerator, washing machine, etc.).

[5] As a sidenote on the implementation of maxentropy in STATA, incorporating too many matching constraints (1) to the maxentropy procedure may result in non-convergence errors and the STATA command may fail to find optimal weights. If so, some matching constraints must be dropped. To minimize the impact on the national averages for the population of interest, it is recommended to drop the matching conditions involving only a limited number of households. For example, if household shares are considered time invariant, they can be matched between the reference and phone surveys. However, including all district level household shares in the matching conditions may result in non-convergence of the algorithm. As a remedy, one can drop the least populous district in each attempt until the maxentropy algorithm converges. This procedure makes the household shares from the most populous districts exactly matched with the reference survey, leaving those in the least populous districts not necessarily matched. However, since the impact of the least populous districts on the national average is minimal, the resulting bias in the national statistics is also minimal.

[6] Since raking can only include discrete variables, continuous variables always need to be discretized before including them in the raking procedure. If raking does not perform well with the discretized variables, such as household size, making the categories finer and re-doing the raking process may yield better results.

Post-Stratification

Post-stratification adjustment matches population/household shares of subnational units between a phone survey and a nationally representative reference survey. If some subnational household shares are largely different from the nationally representative reference survey, even if SWIFT poverty estimation is accurate at the subnational levels, the aggregates or the national average poverty rates could differ largely. PSW cannot fully address these differences-even if subnational level indicators have been included in the PSW procedure, the shares of some subnational units may still exhibit non-negligible differences between the two surveys. Post-stratification, on the other hand, modifies weights so that the population or household share of each subnational unit is exactly matched between the phone survey and the reference survey.

Under the circumstance that maxentropy or raking is executed at the national level and where subnational household or population shares are included as part of the constraints, post-stratification is not necessary. This is because national-level maxentropy or raking matches the subnational shares by including subnational unit indicators. In contrast, if maxentropy or raking is conducted at a subnational level (within each subnational unit), the adjusted household share of each unit will not be matched to that of the reference survey. Therefore, it is important to implement post-stratification after subnational level maxentropy or raking to eliminate any remaining bias due to mismatched population/household shares.

Table 1 Post-Stratification Illustrative Example

Reference Survey Household Distribution		Phone Survey Household Distribution	
District 1	P1	District 1	S1
District 2	P2	District 2	S2
District 3	P3	District 3	S3
District 4	P4	District 4	S4
District 5	P5	District 5	S5
District 6	P6	District 6	S6
District 7	P7	District 7	S7

Table 1 illustrates an example of post-stratification. Suppose our study setting consists of seven districts (District 1 – District 7), with household totals being P1 to P7, respectively, in the reference survey (Panel 1) and S1 to S7 in the phone survey (Panel 2). A district defines a cell in both panels.

A cell weighting or weighting class approach assigns a weight to each cell in the phone sample so that the weighted total within each cell is identical to the reference survey. In list-based or RDD designs, where households (respondents) are selected with simple random sampling, the weighted total is simply the count of observations. In scenarios where initial weights are available, the original weights, together with any weights applied to compensate for subsampling, are carried over to the new survey sample. The sum of these weights is used as the total number of observations within a sample cell.

Let's first assume that the phone survey is a simple random sample. To match the distribution of households across regions between the phone and reference surveys, a factor needs to be applied to all observations of the phone survey sample when producing weights. For example, all observations in the cell of District 1 from the phone survey need to be assigned a weight of (P1/S1), all observations in the cell of District 2 (P2/S2), and the same rule applies to all remaining districts. After this post-stratification exercise, the weighted household distribution across all seven districts became identical to the reference survey. If, in contrast, the phone survey is a subsample of an existing survey with initial weights, then rather than use the factor *per se* as weights, the final weights should be calculated as the product of initial weights and the factor.

This combined approach of (1) PSW, (2) maxentropy/raking, and (3) post-stratification preserves the target variable distributions within each region (cell), meanwhile ensuring that the cross-region (cell) household shares are identical to the reference survey.

III.   Experimental studies on how reweighting procedures and SWIFT poverty projections affect poverty estimates

This section conducts a series of experiments to examine how reweighting procedures and SWIFT modeling affect poverty estimates and other statistics.

First, we create artificially biased subsamples by selecting phone owners from the following reference surveys: Rwanda Integrated Household Living Conditions Survey (EICV) 5, Saint Lucia Household Budget Survey (HBS) 2016, and Uganda Refugee and Host Communities Household Survey (URHS) 2018. These surveys were selected because they are representative of the population of interest in each context and were used as the reference surveys for the pilot of the high-frequency phone surveys implemented by the World Bank (see more details in Yoshida et al., 2022b). Second, we estimate poverty rates among these biased subsamples by applying each of the following methods: 1) selected reweighting procedures alone among the actual consumption data in the phone survey (without SWIFT), 2) SWIFT poverty projection models trained with the reference data, [7] and 3) combinations of SWIFT and reweighting procedures. Finally, we compare each generated poverty rate with the actual poverty rates estimated using the consumption data in the reference survey. A smaller difference suggests better performance of the method in adjusting for the bias in poverty estimates produced using phone owner samples.

We separately examine the following reweighting procedures: (i) original weights; (ii) PSW proposed by Lee (2006), (iii) inverse propensity score approach, and (iv) PSW (Lee 2006) with non-PSW adjustments (maxentropy, raking and/or post-stratification).

Second, we conduct the experiment above using a sample of Ethiopia HFPS7, which was collected based on the ESS4 sample. The subsample is subject to both sampling and nonresponse biases. Using this subsample, we examine whether reweighting procedures and the SWIFT poverty projections combined can reduce sampling and nonresponse biases in poverty estimates.[8]

III.1 Results from experiments with Rwanda EICV 5

Phone ownership is 65.8 percent in Rwanda (9,589 out of 14,574 households own a mobile phone, according to EICV5). As shown in Table 2, the phone owner subsample is significantly wealthier in terms of asset ownership and housing conditions. For example, bicycle ownership among phone owners is 18 percent compared to 13 percent in the full national sample. However, the average household size, which is usually negatively correlated with household welfare, is about 6.8 percent higher in the phone owner subsample than in the full sample (Table 2). While these two effects offset each other, the actual poverty rate of the phone subsample (29.3 percent) is significantly lower than the national average (38.2 percent), indicating that the phone owner subsample is wealthier than the entire population (Figure 2).

Three types of weights were calculated to adjust for the sampling bias in the phone owner sample of Rwanda EICV5, namely, PSW (Lee), PSW (Inverse), and PSW & non-PSW. The first subfigure in Figure 2 compares the poverty rates of the phone owner sample by applying the abovementioned weights to the actual consumption data. The poverty rates among the phone owners applying original weights, PSW (Lee) weights, PSW (Inverse) weights, and PSW & non-PSW weights are 29.3%, 31.4%, 32.0%, and 31.9%, respectively. While the last three poverty rates with adjusted weights are higher than the poverty rate of the phone owner sample

---

[7] In the COVID-19 HFPS for Rwanda, we prepare models for urban and rural areas separately; for Saint Lucia and Uganda (refugees), we prepare only one national SWIFT model.

[8] This analysis is added based on suggestions from peer reviewers of the World Bank's quality enhancement review.

with original weights (29.3%), they are still significantly lower than the national average poverty rate (38.2%). This finding suggests that reweighting alone does not fully address the bias in the phone owner sample.

Figure 2 (subfigure 2) shows that the SWIFT models trained with the full sample of EICV5 project the national poverty rate more accurately. Note that in the case of the Rwanda experiment, we develop urban and rural models separately. Applying the SWIFT models to the reference data set leads to a national projected poverty rate of 37.5 percent,[9] which is within the 95 percent confidence interval of the actual poverty rate of 38.2 percent. The details of the models are available in Appendix 1.

Applying the SWIFT models to the phone owner sample, even with original weights, greatly reduces the sampling bias in poverty estimates. The SWIFT models project a poverty headcount rate of 36.6 percent in the phone owner subsample, which is just 1.6 percentage points lower than the national average rate.

Additional reweighting procedures show slight improvements in poverty projections compared to the abovementioned SWIFT poverty projection with unadjusted weights. PSW and non-PSW adjustments produce a SWIFT-based poverty rate of 37.1 percent, which is around one percentage point smaller than the national average estimate (38.2 percent).[10]

Table 2 displays the summary statistics of all variables involved in this process, namely, variables included in the reweighting procedure, variables included in the SWIFT projection model, and other untargeted indicators not included in either procedure. The weights calculated by PSW-type reweighting techniques produce close estimates for all the aforementioned indicators, and the combination of the PSW and non-PSW adjustments provides incremental improvements.

---

[9] The SWIFT model for all countries can be seen in Appendix 1. The simulation results of the SWIFT model are available upon request.

[10] Adding the non-PSW adjustment (maxentropy), compared to PSW (Lee 2006), has a negligible impact on the SWIFT based poverty projection. The SWIFT-based poverty rate with the inverse propensity score is 37.3 percent, which is the closest to the national poverty rate among all reweighting procedures tested in this experiment. Results are available upon request. Note that the standard errors of all SWIFT-based poverty projections are significantly larger than that of the rural poverty rate estimated from actual consumption data in EICV5 due to the following two reasons: (i) the sample size of the phone owner sample is smaller than the full rural sample of EICV5, and (ii) the SWIFT poverty projections inevitably add prediction errors.

Figure 2 Poverty Projections for the Experiment for Rwanda Round 2

Source: authors' estimation using EICV5 (2017).

Table 2 Summary Statistics for the Experiment in Rwanda

	(1) Reference	(2) Phone (original	(3) Phone (PSW- Lee)	(4) Phone (PSW − inverse)	(5) Phone (PSW & non-PSW)

Figure 2 Poverty Projections for the Experiment for Rwanda Round 2

Source: authors' estimation using EICV5 (2017).

Table 2 Summary Statistics for the Experiment in Rwanda

	(1) Reference	(2) Phone (original	(3) Phone (PSW- Lee)	(4) Phone (PSW − inverse)	(5) Phone (PSW & non-PSW)

Reweighting Target Variables:					
*†Household Size	4.39	4.69	4.46	4.45	4.40
Dependency Ratio	0.43	0.39	0.42	0.42	0.43
*†Household Size Sq.	0.24	0.27	0.24	0.24	0.24
Urban/Rural	0.19	0.26	0.20	0.20	0.18
Head Age	44.70	43.25	44.44	44.42	44.79
†region==Kigali	0.15	0.20	0.16	0.15	0.14
region==Other urban	0.08	0.10	0.09	0.09	0.08
region==Southern rural	0.21	0.18	0.21	0.21	0.21
*region==Western rural	0.18	0.17	0.18	0.18	0.19
*region==Northern rural	0.14	0.13	0.13	0.14	0.14
wall==mud/bricks/cement	0.69	0.74	0.69	0.69	0.69
wall==wood/stone/tree trunks/plastic/other	0.31	0.26	0.31	0.31	0.31
lighting==Elec/biogas/generator	0.34	0.45	0.35	0.35	0.33
lighting==oil lamp/firewood/candle/ lantern	0.09	0.12	0.12	0.13	0.12
lighting==solar/batteries/phone	0.56	0.43	0.53	0.53	0.54
TV set	0.10	0.15	0.11	0.11	0.11
Bicycle	0.13	0.18	0.14	0.14	0.14
Decoder	0.07	0.11	0.08	0.08	0.07
Cupboard	0.10	0.15	0.12	0.11	0.11
Additional SWIFT model variables (rural)					
†Head Sex	0.25	0.20	0.24	0.24	0.25
head age (square/100)	22.41	20.73	22.09	22.11	22.54
head_mrstat==Widow or widower	0.17	0.13	0.16	0.16	0.17
†Worked at least 1h in last 7 days	0.86	0.88	0.87	0.87	0.87
†head_sec==2 Private farm	0.53	0.47	0.51	0.51	0.52
consume local/imported rice in the past 7 days	0.37	0.42	0.40	0.40	0.40
Additional SWIFT model Variables (urban)					
Head Age	44.70	43.25	44.44	44.42	44.79
consume beef meat in the past 7 days	0.11	0.15	0.13	0.12	0.12
Others:					
head_mrstat==Married monogamously	0.62	0.67	0.63	0.64	0.62
Radio	0.45	0.56	0.53	0.53	0.53
Observations	14580	9589	9589	9589	9589

Note: Variables marked with * are used both as reweighting target variables and as variables included in the SWIFT poverty projection model. To avoid repetition, we only include these variables in the first section (reweighting target variables) or in the first section where they show up (SWIFT model Variables-rural) but exclude them from subsequent sections (SWIFT model variables-urban).
Source: authors' estimation using EICV5 (2017).

III.2 Results from experiments with Saint Lucia HBS 2016 data

For Saint Lucia's experiment, the poverty estimate according to the official statistics is 25.0 percent, and the poverty rate among the subsample without any weight adjustment is 22.2 percent (Figure 3). The difference of 2.8 percentage points between the actual rates of the reference and phone surveys is not statistically significant at the five percent level. This small difference stems from widespread mobile phone ownership; in the Household Budget Survey 2016, 81.5 percent of households (1,214 out of 1,490) stated they own a mobile phone.

Like Rwanda EICV5 in the previous section, three types of weights were calculated to adjust for the sampling bias in the phone owners' sample of HBS 2016 in Saint Lucia, namely, PSW (Lee), PSW (Inverse), and PSW & non-PSW. The first subfigure in Figure 3 compares the poverty rates of the phone owner sample by applying the abovementioned weights to the consumption data. The poverty rates among the phone owners applying original weights, PSW (Lee) weights, PSW (Inverse) weights, and PSW & non-PSW weights are 22.2%, 23.5%, 24.2%, and 24.7%, respectively. Among the last three poverty rates with adjusted weights, the performance of

the combination of PSW and non-PSW weighting is the best (24.7% versus 25.0% in the reference sample). The combination virtually greatly reduces the bias.

Figure 3 (subfigure 2) shows that the SWIFT model trained with the full sample of the reference data (including both phone owners and non-phone owners) projects the national poverty rate more accurately. Note that in the case of the Saint Lucia experiment, we develop only one model. The model projects a national poverty rate of 25.3 percent, which is within the 95 percent confidence interval of the actual poverty rate of 25.0 percent (the simulation results are available upon request). The details of the model are available in Appendix 1.

Applying the SWIFT model to the phone owner sample reduces the sampling bias in poverty estimation. Using original weights, the SWIFT model projects a national poverty rate of 24.4 percent, which reduces the bias from the actual national average poverty rate (25.0 percent) from the previous 2.8 percentage points to the current 0.6 percentage points.

All reweighting procedures show further but minor improvements in poverty projections. PSW with a non-PSW adjustment (maxentropy) produces a national poverty rate of 26.0 percent, which is 1.0 percentage points apart from the national average estimate (25.0 percent). The standard errors of all SWIFT-based poverty projections are larger than that of the national poverty rate directly estimated from the actual consumption data (displayed as "Reference (direct)" in Figure 3) because of the relatively small number of observations in the phone owner sample and the resulting lack of statistical power.

Lastly, Table 3 displays the summary statistics of variables included in the reweighting procedure, variables included in the SWIFT projection model, and other untargeted indicators not included in either procedure. All reweighting techniques produce close estimates for all the above-mentioned indicators, and the difference between reweighting procedures is minimal.

Figure 3 Poverty Projections for the Experiment in Saint Lucia
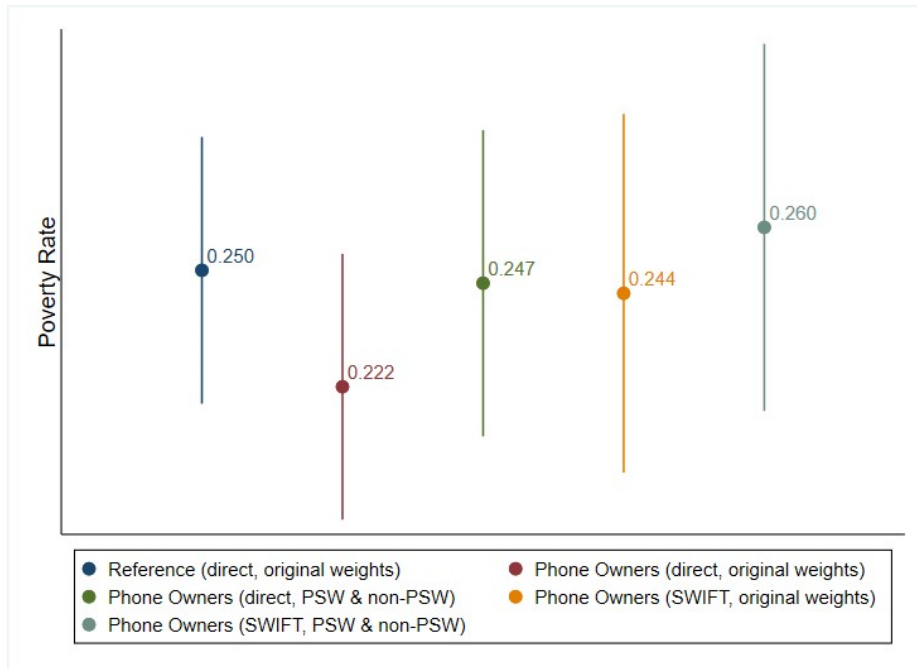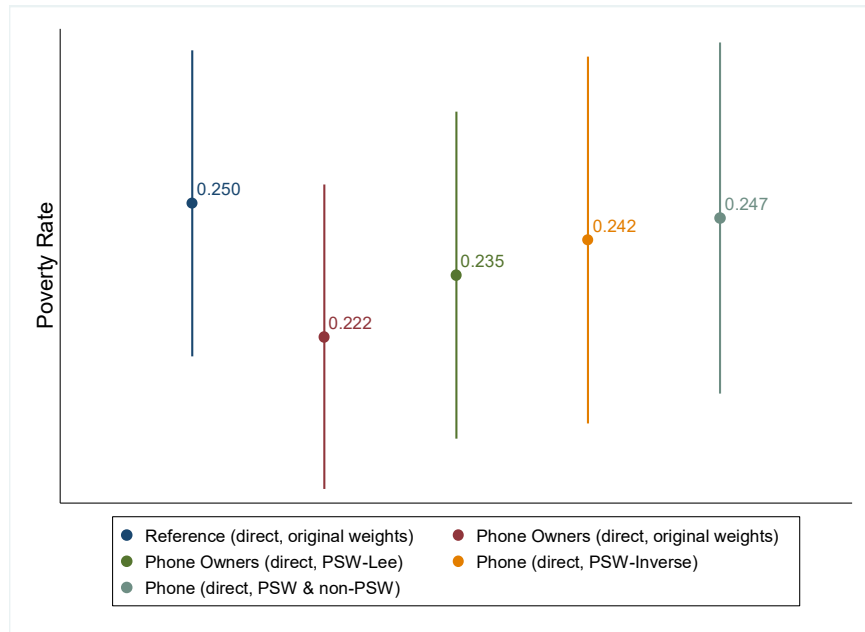
Source: authors' estimation using HBS 2016.

Table 3 Summary Statistics for the experiment for Saint. Lucia

	(1) Reference	(2) Phone (original weights)	(3) Phone (PSW-Lee)	(4) Phone (PSW – inverse)	(5) Phone (PSW & non-PSW)
Reweighting Target Variables:					
*Urban/Rural	0.70	0.72	0.70	0.70	0.70
*Household Size	3.07	3.15	3.07	3.09	3.07
*Dependency Ratio	0.33	0.30	0.32	0.32	0.33

*Household Size Sq.	0.13	0.14	0.13	0.14	0.13
*Sex of Head (1=male)	0.60	0.60	0.60	0.60	0.60
Internet access installed at home	0.38	0.43	0.40	0.38	0.39
Wall==plywood	0.17	0.16	0.17	0.17	0.17
Toilet==flush toilet	0.75	0.78	0.77	0.77	0.75
Land_owner==owned with title	0.29	0.29	0.29	0.28	0.29
Smart television/television	0.82	0.85	0.84	0.84	0.82
*District==Castries City	0.15	0.15	0.15	0.15	0.15
*District==Castries Sub-Urban	0.25	0.27	0.26	0.26	0.25
*District==AnseLaRayeCanaries	0.05	0.04	0.04	0.05	0.05
District==Soufriere	0.05	0.05	0.05	0.05	0.05
District==Laborie	0.04	0.03	0.04	0.04	0.04
*District==Vieuxfort	0.09	0.08	0.08	0.08	0.09
District==Micoud	0.10	0.10	0.10	0.10	0.10
District==Dennery	0.07	0.07	0.08	0.08	0.07
District==Gros Islet	0.15	0.16	0.15	0.15	0.15
Additional SWIFT model Variables					
Age of Head	52.26	49.85	50.51	50.44	50.70
you went without eating for a whole day	0.07	0.05	0.05	0.05	0.06
your household ran out of food because of a lack of money	0.20	0.18	0.18	0.18	0.18
Employment Status	0.65	0.72	0.70	0.70	0.70
Others:					
Wall==wood/timber	0.13	0.11	0.11	0.11	0.12
Wall==blocks	0.51	0.54	0.53	0.52	0.52
Wall==wood/concrete	0.18	0.18	0.18	0.19	0.18
Number of rooms	3.57	3.63	3.61	3.60	3.60
Refrigerator	0.81	0.85	0.84	0.84	0.83
Freezer	0.13	0.15	0.14	0.14	0.14
Washing machine	0.56	0.62	0.60	0.59	0.58
Smart devices – tablet	0.21	0.25	0.23	0.23	0.23
Jewellery	0.14	0.16	0.16	0.15	0.15
Observations	1491	1214	1214	1213	1213

Note: Variables marked with asterisks are used both as reweighting target variables and as variables included in the SWIFT poverty projection model. To avoid repetition, we only include these variables in the first section (reweighting target variables) but exclude them from the second section (SWIFT model variables).
Source: authors' estimation using HBS 2016.

III.3. Results from experiments with the 2018 URHS

The poverty estimate of the 2018 URHS is 47.1 percent, and the actual poverty headcount rate of the phone owner subsample is 38.8 percent. The lower poverty rate with the unadjusted weights in the phone sample resonates with the findings in the existing literature that the phone owners are much richer than the full sample of refugees.

Like Rwanda EICV5 and Saint Lucia HBS 2016, three types of weights were calculated to adjust for the sampling bias in the phone owner sample of URHS 2018 in Uganda, namely, PSW (Lee), PSW (Inverse), and PSW & non-PSW.[11] The first subfigure in Figure 4 compares the poverty rates of the phone owner sample by applying the abovementioned weights to the consumption data. The poverty rates among the phone owners applying original weights, PSW (Lee) weights, PSW (Inverse) weights, and PSW & non-PSW weights are 38.8%, 43.6%, 43.0%, and 42.6%, respectively. While the last three poverty rates with adjusted weights are slightly but not significantly higher than the poverty rate of the phone owner sample with original weights (38.8%), they

---

[11] External data sources on the share of refugees from each country of origin are used to calibrate the weights. Details of the non-PSW adjustments are available in Appendix 3.

still seem to be lower than the poverty rate of the full sample (47.1%). This finding suggests that reweighting alone does not fully address the bias in the phone owner sample.

Next, we apply a SWIFT-based poverty projection model and reweighting techniques to the biased phone owner sample. (The simulation results of the SWIFT model are available upon request. The details of the model are available in Appendix 1.) As can be seen from Figure 4, applying the SWIFT model to the phone owner sample without reweighting greatly reduces the sampling bias in poverty estimation. The SWIFT model projects a poverty rate of 44.4 percent for the full sample of refugees with original weights, which is only 2.7 percentage points lower than the actual poverty headcount rate among all refugees (47.1 percent).

All reweighting procedures show further improvements in SWIFT poverty projections. PSW and non-PSW reweighting approaches produce a poverty rate of 46.5 percent, which is the closest to the refugee average poverty rate among all reweighting procedures.[12]

Table 4 displays the summary statistics of variables included in the reweighting procedure, variables included in the SWIFT projection model, and other untargeted indicators not included in either procedure. All reweighting techniques produce close estimates for all the above-mentioned indicators, and the difference between reweighting procedures is minimal.

---

[12] PSW (Lee 2006) and PSW with non-PSW adjustments produce SWIFT based poverty rates of 46.7 and 46.5 percent, respectively. Both estimates are within one percentage point from the actual poverty estimate (47.1 percent). Results are available upon request.

Figure 4 Poverty Projections for the Experiment in Uganda Refugees
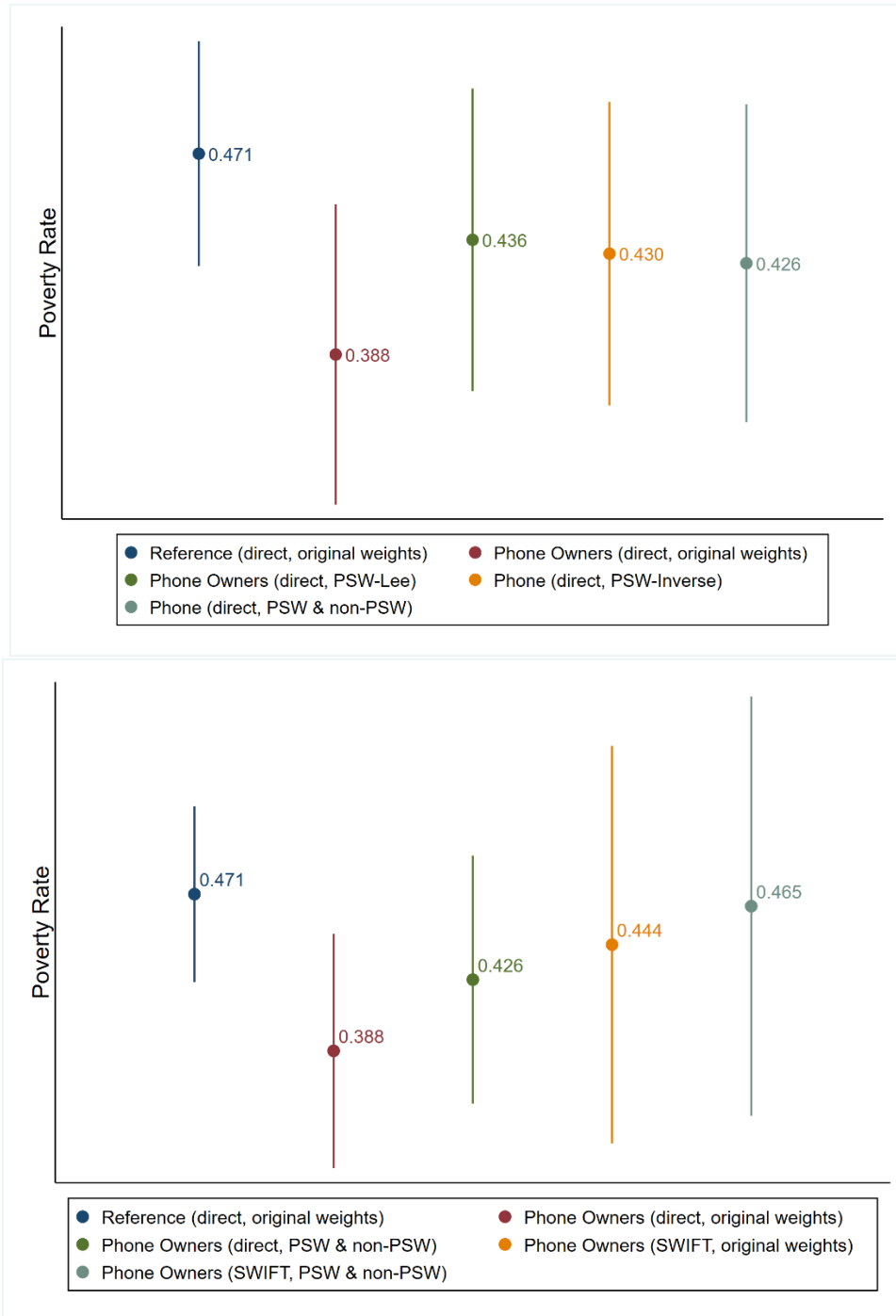
Source: authors' estimation using URHS 2018.

Table 4 Summary Statistics for the Experiment in Uganda Refugees

	(1) Reference	(2) Phone (original weights)	(3) Phone (PSW-Lee)	(4) Phone (PSW – inverse)	(5) Phone (PSW & non-PSW)
Reweighting Target Variables:					
*Household Size	5.53	5.98	5.58	5.55	5.53
Dependent% of residents	0.55	0.51	0.54	0.54	0.55
*HH Size Squared	38.99	44.56	39.14	39.23	38.99
REGION==Kampala	0.05	0.09	0.06	0.05	0.05
*REGION==West Nile	0.66	0.59	0.67	0.67	0.66
*REGION==Southwest	0.29	0.32	0.27	0.28	0.29
Asset: Furniture/Furnishings	0.43	0.55	0.42	0.43	0.41
Asset: Motor cycle	0.02	0.03	0.02	0.02	0.02
Asset: Bicycle	0.07	0.09	0.07	0.07	0.06
HH has more than 1 room	0.55	0.64	0.55	0.56	0.56
Age of Head	39.31	38.49	38.08	38.47	39.31
Gender of Head (Male)	0.46	0.56	0.52	0.51	0.46
Additional SWIFT model Variables					
Arrived before 2008	0.05	0.05	0.04	0.04	0.04
Arrived between 2008 and 2015	0.27	0.34	0.28	0.29	0.30
Income Source 12M: Remittances	0.28	0.34	0.28	0.29	0.31
Scores for pca for asset ownership	0.73	0.67	0.69	0.69	0.70
Others:					
#Sleeping Rooms	1.81	1.98	1.81	1.81	1.83
income source=wage and salary	0.21	0.24	0.23	0.22	0.22
Consumption: rice	0.14	0.20	0.16	0.16	0.16
Toilet_type==Flush toilet	0.03	0.06	0.04	0.03	0.03
Country of Origin==South Sudan	0.72	0.66	0.73	0.73	0.72
Country of Origin==DRC	0.23	0.25	0.20	0.21	0.22
Country of Origin==Burundi	0.03	0.03	0.03	0.03	0.03
Country of Origin==Somalia	0.03	0.06	0.04	0.03	0.03
Observations	806	437	437	437	437

Note: Variables marked with asterisks are used both as reweighting target variables and as variables included in the SWIFT poverty projection model. To avoid repetition, we only include these variables in the first section (reweighting target variables) but exclude them from the second section (SWIFT model variables).

Source: authors' estimation using URHS 2018.

III.4. Results from experiments with the Ethiopia ESS round 4 and HFPS round 7 data

Background of the Ethiopia ESS and HFPS data and creation of a biased subsample

The Ethiopia HFPS monitors the economic and social impacts of the COVID-19 pandemic on households by interviewing a sample of households over 15 months for twelve survey rounds.

The HFPS sample is a subsample of the 2018/19 Ethiopia Socioeconomic Survey round 4 (ESS4). The ESS collects panel data on household and community characteristics in both rural and urban areas. Four waves have been conducted since 2011, and ESS4 is the most recent in 2018/19. ESS4 included a total of 6,770 households. In the ESS4 interview, households were asked to provide phone numbers, either of their own or that of a reference household (i.e., a friend's or neighbor's), so they can be contacted in the follow-up ESS surveys should they move from their sampled location. At least one valid phone number was obtained for 5,374 households. These households established the sampling frame for the HFPS. While the Ethiopia COVID-19 HFPS drew its sample from the ESS4 database, the final sample size decreased to 3,249 households due to non-responses.

The phone penetration rate in rural Ethiopia is around 40 percent, which contrasts with a phone penetration rate of over 90 percent in urban Ethiopia. The ESS4 data was used as a sampling frame, a reference survey for reweighting, and for developing the SWIFT-based poverty projection model for Ethiopia's pre-COVID poverty estimates.

Ethiopia conducted twelve rounds of HFPS between April 2020 and June 2021. Round-on-round attrition was high due to (i) enumerator and respondent fatigue, (ii) challenges with network connectivity, particularly in rural Ethiopia; and (iii) a conflict that erupted in the North of Ethiopia and that prevented calls among households in the Tigray regions after the seventh round. Round seven, which included a SWIFT module, was conducted between October 19 and November 10, 2020, and interviewed 2,534 households (715 rural and 1,819 urban). HFPS round seven (HFPS7) data is the main data set used for poverty and inequality analysis and profiling of the poor.

The sample of the HFPS7, similar to all other rounds of the Ethiopia HFPS, is subject to both sampling bias and non-responses bias. The selection of phone owners from the ESS4 reference survey leads to sampling bias, and the random digit dial process as part of the HFPS4 further leads to nonresponse bias. Therefore, by construction, this subsample is subject to both types of biases.
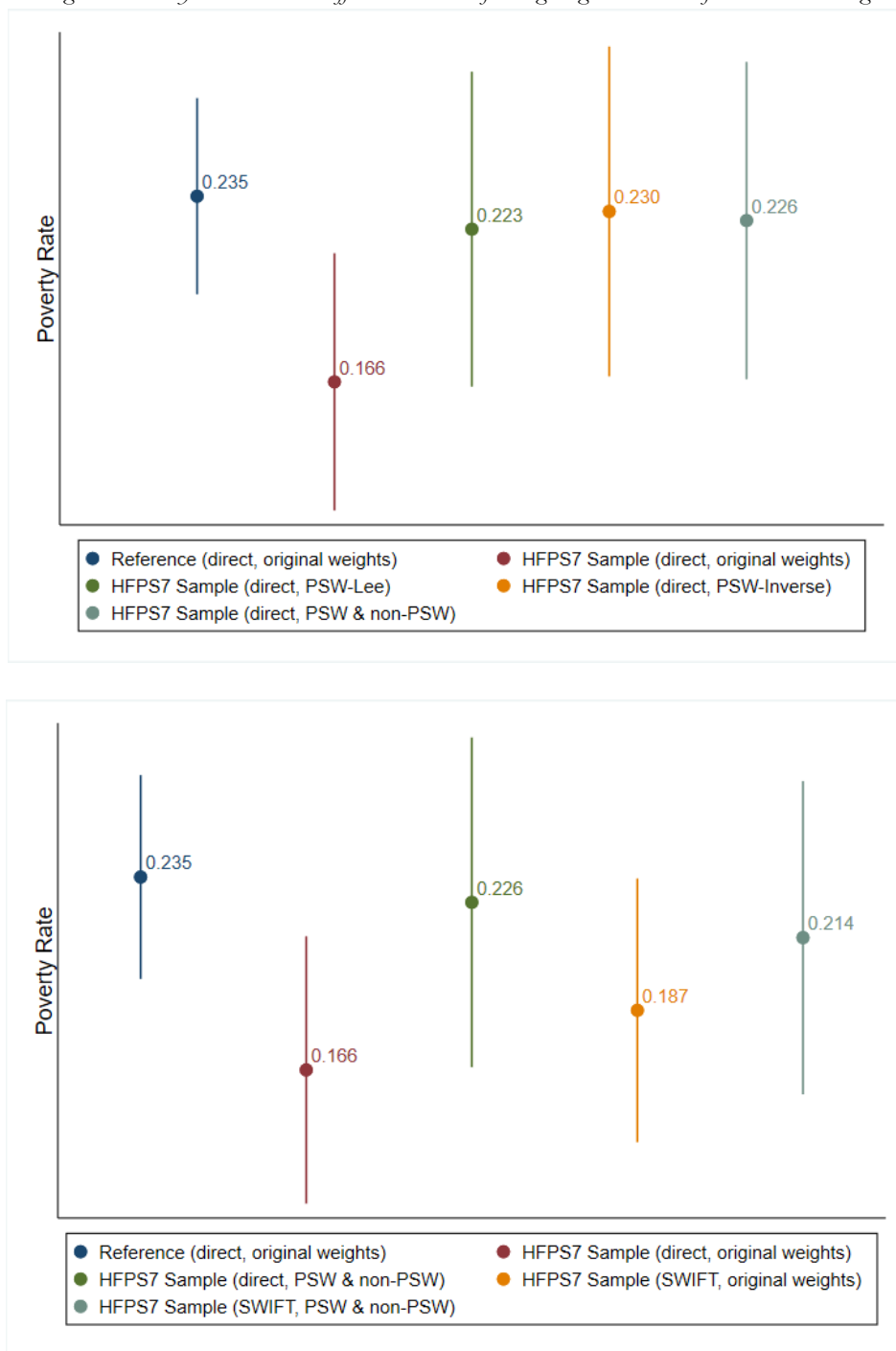
Results

In this Ethiopia example, the poverty estimate from the reference sample is 23.5 percent, and the poverty rate among the subsample of phone owners without any weight adjustment is 16.6 percent (Figure 5). The difference of approximately seven percentage points between the actual rate of the reference survey and that of the phone survey is not statistically significant at the five percent level.

Like previous experiments, three types of weights were calculated to adjust for the sampling bias in the subsample of ESS4, namely, PSW (Lee), PSW (Inverse), and PSW & non-PSW. The first subfigure in Figure 5 compares the poverty rates of the phone owner sample by applying the abovementioned weights to the consumption data. The poverty rates among the biased HFPS subsample applying original weights, PSW (Lee) weights, PSW (Inverse) weights, and PSW & non-PSW weights are 16.6%, 22.3%, 23.0%, and 22.6%, respectively. These results show that different types of reweighting procedures reduce the bias due to sampling and nonresponses but do not fully eliminate the bias.

Figure 5 (subfigure 2) shows that the SWIFT model trained with the full sample of the reference data, when combined with reweighting techniques, accurately projects the national poverty rate, but does not show almost any clear improvement compared with the results with only reweighting. The SWIFT-based poverty rate using original weights is 18.7%, and the SWIFT-based poverty rate applying PSW and non-PSW weights is 21.4%. These rates are not as close to the poverty rate in the reference survey (23.5%) as the poverty rate estimates using reweighting procedures. This suggests that once reweighting alleviates sampling and nonresponse bias, the SWIFT projection method does not make additional improvements.

Lastly, Table 5 displays the summary statistics of variables included in the reweighting procedure, variables included in the SWIFT projection model, and other untargeted indicators not included in either procedure. All reweighting techniques produce close estimates for all the above-mentioned indicators, and the difference between reweighting procedures is minimal.

Source: authors' estimation using ESS4 and HFPS7.

	(1) Reference	(2) HFPS HH (original weights)	(3) HFPS HH(PSW-Lee)	(4) HFPS HH (PSW-Inverse)	(5) HFPS HH(PSW-non-PSW)
Reweighting PSW Target Variables:					
National					
Head age	42.91	41.20	42.08	42.26	43.12
Head age sq.	2076.81	1905.34	1987.01	2004.68	2084.89
Head male	0.74	0.75	2.29	2.30	2.23
Head marital status = married	0.74	0.74	0.76	0.76	0.75
Head education: incomplete primary school	0.25	0.27	0.24	0.25	0.25
Head education: complete primary school	0.04	0.06	0.05	0.05	0.05
Head education: incomplete secondary school	0.08	0.11	0.08	0.08	0.07
Head education: complete secondary school	0.03	0.04	0.03	0.02	0.03
Head education: post-secondary school	0.08	0.14	0.09	0.08	0.09
Head labor market status: unemployed	0.05	0.06	0.05	0.04	0.04
Head labor market status: inactive	0.19	0.18	0.17	0.17	0.18
Head engaged in agriculture	0.55	0.44	0.55	0.57	0.55
Head engaged in casual labor activities	0.04	0.04	0.03	0.03	0.03
Received remittances	0.30	0.30	0.29	0.29	0.32
House ownership	0.77	0.68	0.77	0.78	0.77
Number of people per room	0.52	0.56	0.52	0.52	0.54
Urban/rural status	0.32	0.47	0.33	0.31	0.32
Consumption quintile = 2	0.20	0.17	0.20	0.20	0.20
Consumption quintile = 3	0.20	0.24	0.22	0.22	0.21
Consumption quintile = 4	0.20	0.23	0.21	0.21	0.21
Consumption quintile = 5	0.20	0.25	0.20	0.20	0.20
Household size	4.68	4.63	4.88	4.89	4.73
Reweighting non-PSW Target Variables:					
Urban					
Household size	3.65	3.70	3.71	3.70	3.71
Head age	38.73	38.80	39.06	39.14	38.97
Head education: incomplete primary school	0.23	0.23	0.22	0.23	0.22
Head education: complete primary school	0.07	0.08	0.07	0.07	0.07
Head education: incomplete secondary school	0.14	0.15	0.14	0.14	0.14
Head education: complete secondary school	0.07	0.08	0.07	0.06	0.07
Head education: post-secondary school	0.20	0.24	0.20	0.18	0.21
Rural					
Household size	5.18	5.46	5.46	5.43	5.22
Head age	44.92	43.36	43.59	43.66	45.10
Head education: incomplete primary school	0.26	0.30	0.25	0.25	0.26
Head education: complete primary school	0.03	0.04	0.03	0.03	0.03

Head education: incomplete secondary school	0.04	0.08	0.05	0.06	0.04
Head education: complete secondary school	0.01	0.00	0.00	0.00	0.01
Head education: post-secondary school	0.02	0.05	0.03	0.03	0.03
SWIFT model Variables:					
Urban					
Region==SNNP	0.15	0.11	0.11	0.11	0.15
Reduced Consumption Preferred Foods Past 7 Days	0.21	0.19	0.21	0.22	0.21
Reduce number of meals eaten in a day	0.16	0.13	0.15	0.16	0.15
Anyone in HHs is in agriculture last week	0.15	0.14	0.16	0.17	0.15
Anyone in HHs is in non-farm business last week	0.26	0.27	0.26	0.25	0.25
Household Size	3.65	3.70	3.71	3.70	3.71
Household size squared	1.75	1.75	1.77	1.76	1.80
Head can read and write	0.75	0.82	0.76	0.74	0.76
Number of rooms(excluding kitchen, toilet and bath	1.97	2.08	2.06	2.03	2.04
Floor==cement creed	0.42	0.44	0.41	0.39	0.41
Floor==tiles	0.06	0.07	0.06	0.06	0.06
Floor==other	0.01	0.01	0.01	0.01	0.01
Rural					
Region==TIGRAY	0.07	0.07	0.06	0.06	0.07
Region==AMHARA	0.27	0.25	0.27	0.27	0.26
Region==SNNP	0.21	0.10	0.09	0.09	0.20
Reduced Consumption Preferred Foods Past 7 Days	0.24	0.22	0.22	0.22	0.24
Reduce number of meals eaten in a day	0.17	0.14	0.15	0.15	0.16
Anyone in HHs is in agriculture last week	0.85	0.83	0.84	0.84	0.82
Anyone in HHs is in non-farm business last week	0.08	0.10	0.09	0.09	0.10
Household Size	4.88	5.07	5.07	5.03	4.87
Household size squared	2.86	3.07	3.07	3.02	2.86
Number of rooms(excluding kitchen, toilet and bath	1.85	2.05	2.02	2.03	2.04
Floor==cement creed	0.02	0.05	0.04	0.04	0.04
Floor==tiles	0.01	0.01	0.01	0.01	0.01
Floor==other	0.02	0.02	0.02	0.02	0.02
Other variables:					
Urban					
Consumed wheat	0.32	0.34	0.33	0.32	0.32
Consumed pasta/maccaroni	0.35	0.38	0.35	0.34	0.36
Consumed tomato	0.61	0.64	0.60	0.58	0.60
Consumed sugar	0.73	0.76	0.73	0.72	0.73
Consumed Lentils	0.43	0.49	0.46	0.44	0.44
Consumed banana	0.33	0.35	0.34	0.32	0.35
Consumed milk	0.24	0.23	0.21	0.20	0.23
Consumed tea	0.54	0.56	0.54	0.53	0.54
Consumed green chili pepper	0.68	0.72	0.71	0.70	0.69

25

Consumed potato	0.74	0.76	0.75	0.74	0.74
Consumed coffee	0.78	0.83	0.83	0.83	0.80
Household bought candles last month	0.46	0.52	0.49	0.48	0.48
Household bought batteries last month	0.29	0.25	0.27	0.27	0.28
Household bought laundry soap last month	0.88	0.89	0.89	0.89	0.88
Household bought kerosene last month	0.10	0.11	0.12	0.12	0.12
Toilet==flush	0.27	0.29	0.27	0.26	0.29
Toilet==pit latrine	0.65	0.65	0.67	0.67	0.65
Toilet==no facility	0.08	0.05	0.06	0.07	0.05
Toilet==other	0.00	0.00	0.00	0.00	0.00
Lighting=electricity meter	0.87	0.91	0.90	0.89	0.89
Lighting=solar energy	0.03	0.03	0.03	0.04	0.03
Lighting=flashlight	0.02	0.01	0.01	0.01	0.01
Lighting=else	0.08	0.05	0.06	0.06	0.06
Rural					
Consumed wheat	0.32	0.39	0.37	0.37	0.35
Consumed pasta/maccaroni	0.11	0.14	0.13	0.13	0.14
Consumed tomato	0.22	0.23	0.21	0.21	0.21
Consumed sugar	0.43	0.53	0.52	0.52	0.50
Consumed Lentils	0.09	0.12	0.10	0.10	0.10
Consumed banana	0.14	0.14	0.13	0.13	0.15
Consumed milk	0.28	0.32	0.31	0.31	0.33
Consumed tea	0.17	0.17	0.16	0.16	0.18
Consumed green chili pepper	0.43	0.49	0.46	0.47	0.47
Consumed potato	0.45	0.51	0.49	0.50	0.49
Consumed coffee	0.80	0.86	0.86	0.86	0.84
Household bought candles last month	0.09	0.11	0.10	0.10	0.09
Household bought batteries last month	0.65	0.63	0.61	0.62	0.63
Household bought laundry soap last month	0.88	0.94	0.93	0.93	0.92
Household bought kerosene last month	0.24	0.24	0.24	0.24	0.24
Toilet==flush	0.02	0.01	0.01	0.01	0.01
Toilet==pit latrine	0.54	0.60	0.58	0.58	0.60
Toilet==no facility	0.43	0.38	0.40	0.40	0.39
Toilet==other	0.01	0.01	0.01	0.01	0.01
Lighting=electricity meter	0.09	0.14	0.11	0.12	0.12
Lighting=solar energy	0.30	0.38	0.39	0.39	0.40
Lighting=flashlight	0.19	0.15	0.15	0.15	0.14
Lighting=else	0.42	0.33	0.35	0.35	0.35
Number of Observations:					
Urban	3655	1819	1819	1819	1819
Rural	3115	715	715	715	715

Source: authors' estimation using ESS4 and HFPS7.

IV.   Concluding remarks

Reweighting procedures do not fully correct the bias in poverty estimation, and the selection of reweighting procedures listed above does not make much difference in correcting the bias. However, our findings show that a combination of reweighting and the SWIFT-based poverty projection approach effectively eliminates the bias in poverty estimation in all four case studies. The literature proposes several weight adjustment procedures to reduce sampling bias in phone or web surveys. However, as far as we know, there has been no previous attempt that examined the role of reweighting on poverty projections.

To monitor the impacts of the COVID-19 pandemic in a frequent and timely manner, the World Bank launched the COVID-19 HFPS surveys in around 80 countries. The COVID-19 HFPS surveys offer valuable and timely information on the living conditions, livelihoods, coping mechanisms, and social assistance during the pandemic, but they also have serious limitations. First, these surveys do not have direct poverty or inequality measures. Second, the data collected often suffers from sampling and nonresponse bias due to limited phone ownership in many developing countries and nonresponses associated with phone interviews. To address the first limitation, SWIFT methodology was adopted. SWIFT utilizes a machine-learning-based technique to impute poverty and inequality statistics using 10 to 15 simple questions included in the HFPS surveys (Yoshida et al., 2022a). However, the second limitation, namely, bias arising from sampling and nonresponses, can lead to bias in the poverty statistics estimated by SWIFT. To address this issue, this paper explores some widely used reweighting techniques and examines if they could alleviate these bias in poverty projections.

Specifically, this paper identifies a few variants of propensity score weightings (PSWs), including 1) PSWs and 2) a combination of PSW and non-PSW techniques. To assess the performance of reweighting procedures and the SWIFT-based poverty projections, we conduct experiments by drawing subsamples of phone owners from reference surveys in Rwanda, St Lucia, Uganda (refugees), and Ethiopia. In the experiment in Ethiopia, we excluded households who did not own a phone from the ESS4 survey and further excluded those who did not respond to the Ethiopia High-Frequency Phone Survey, which leads to both sampling bias and nonresponse bias by construction. Using these artificially biased subsamples, we test whether reweighting procedures and the SWIFT poverty projections combined can reduce the sampling and nonresponse bias in poverty estimation.

Interestingly, the contribution of reweighting procedures and the SWIFT-based poverty projection technique in reducing the bias of the poverty estimates differ by context. In the experiments of Rwanda EICV5 and Uganda Refugee Household Survey, the SWIFT poverty projections are the main contributor to reducing the bias in poverty estimation. In the experiments with Ethiopia ESS4 and St Lucia 2016 HBS surveys, reweighting procedures are the main contributor to reducing the bias in poverty estimation.

How does the SWIFT approach to poverty projection play a role in minimizing sampling bias? Existing literature indicates that while reweighting can mitigate sampling bias, it cannot completely remove it. This paper's experiments also confirm that poverty rates derived from actual consumption data are still subject to bias. In contrast, the poverty rates imputed by SWIFT, post-reweighting, have almost no sampling bias in all four examined cases. What accounts for this difference?

A plausible explanation lies in the limitations of reweighting. Even though it balances many observed variables between a biased subsample and the full sample, reweighting may not balance certain unobserved variables correlated with household expenditures and poverty. Consequently, poverty rates based on actual consumption data, which are correlated with the unobserved but unbalanced variables, could retain the sampling bias. In contrast, SWIFT's poverty projections are not influenced by these unobserved household characteristics in imputing household-level consumption. Thus, if all regression variables in the SWIFT models are effectively balanced through reweighting, the resulting poverty rates should be unbiased. However, further theoretical and empirical studies are essential for definitive conclusions. `

References

Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46: 399–424.

Capacci, Sara, Mario Mazzocchi, and Sergio Brasini. 2018. "Estimation of unobservable selection effects in on-line surveys through propensity score matching: An application to public acceptance of healthy eating policies." PLoS ONE 13(4): e0196020. https://doi.org/10.1371/journal.pone.0196020.

Central Statistics Agency of Ethiopia. 2019. "Socioeconomic Survey 2018-2019, Round 4 (ESS4)." https://doi.org/10.48529/k739-c548. (Reference No.: ETH_2018_ESS_v03_M).

Central Statistical Office of Saint Lucia. 2016. "Survey of Living Conditions and Household Budgets (HBS)." Ministry of Finance, Economic Affairs and Social Security, Government of Saint Lucia. (Reference No.: LCA_2015_SLC-HBS_v01_M).

Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics*: 295–313.

Czajka, John L., Sharon M. Hirabayashi, Roderick J. A. Little, and Donald B. Rubin. 1992. "Projecting From Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics." *Journal of Business & Economic Statistics* 10 (2): 117–31. https://doi.org/10.1080/07350015.1992.10509892.

Drake, Christiana. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." *Biometrics* 49 (4): 1231–36. https://doi.org/10.2307/2532266.

Dreze, J. and A. Somanchi, 2023. "Weighty evidence? Poverty estimation with missing data." Ideas for India. https://www.ideasforindia.in/topics/poverty-inequality/weighty-evidence-poverty-estimation-with-missing-data.html

Duncan, K. B., and E. A. Stasny. 2001. "Using Propensity Scores to Control Coverage Bias in Telephone Surveys." *Survey Methodology* 27: 121–30.

Golan, Amos, George G. Judge, and Douglas Miller. 1996. "Maximum Entropy Econometrics." Iowa State University, Department of Economics.

Jaynes, Edwin T. 1957. "Information Theory and Statistical Mechanics." *Physical Review* 106 (4): 620.

Kolenikov S. Calibrating Survey Data using Iterative Proportional Fitting (Raking). The Stata Journal. 2014;14(1):22-59. doi:10.1177/1536867X1401400104

Lee, Sunghee. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys." *Journal of Official Statistics* 22 (2): 329–49.

Li, Fan, Kari Lock Morgan, and Alan M. Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400. https://doi.org/10.1080/01621459.2016.1260466.

Little, Roderick J., and Sonya Vartivarian. 2003. "On Weighting the Rates in Nonresponse Weights." *Statistics in Medicine* 22 (9): 1589–99. https://doi.org/10.1002/sim.1513.

Morgan, Stephen L. and Jennifer J. Todd. 2008. "6. A Diagnostic Routine for the Detection of Consequential Heterogeneity of Causal Effects." Sociological Methodology 38 (1): 231-282.

National Institute of Statistics Rwanda (NISR) - Ministry of Finance and Economic Planning. 2017. "Integrated Household Living Conditions Survey 5 (EICV 5)." (Reference No.: RWA-NISR-EICV5-CS-2016-2017-V0.1).

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55. https://doi.org/10.1093/biomet/70.1.41.

———. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–24. https://doi.org/10.1080/01621459.1984.10478078.

Rubin, Donald.B. and Neal Thomas. 1996. "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics* 52: 249-264. http://dx.doi.org/10.2307/2533160

Schafer, Joseph L., Joseph Kang. 2008. "Average Causal Effects From Nonrandomized Studies: A Practical

Guide and Simulated Example." Psychological Methods 13 (4): 279–313.

Smith, Philip J, J N K Rao, Danni Daniels, Trena Ezzati-Rice, and Meena Khare. 2000. "Compensating For Nonresponse Bias In The National Immunization Survey Using Response Propensities." *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 641–46.

Taylor, Humphrey. 2000. "Does Internet Research Work? Comparing Online Survey Result with Telephone Survey." *International Journal of Market Research* 42 (1): 58–63. https://doi.org/10.1177/147078530004200104.

Taylor, Humphrey, John Bremer, Cary Overmeyer, Jonathan W. Siegal, and George Terhanian. 2001. "The Record of Internet-Based Opinion Polls in Predicting the Results of 72 Races in the November 2000 US Elections." *International Journal of Market Research* 43 (2): 1–10. https://doi.org/10.1177/147078530104300203.

Terhanian, George, John Bremer, Renee Smith, and Randy Thomas. 2000. "Correcting Data from Online Surveys for the Effects of Nonrandom Selection and Nonrandom Assignment." *Harris Interactive*.

The World Bank. 2018. "Refugee and Host Communities Household Survey 2018 (URHS 2018)." https://doi.org/10.48529/gvwc-vx89. (Reference No.: UGA_2018_RHCS_v01_M)

Wittenberg, Martin. 2010. "An Introduction to Maximum Entropy and Minimum Cross-Entropy Estimation Using Stata." *The Stata Journal: Promoting Communications on Statistics and Stata* 10 (3): 315–30. https://doi.org/10.1177/1536867X1001000301.

Yoshida, N., X. Chen, S. Takamatsu, K. Yoshimura, S. Malgioglio, S. Shivakumaran, K. Zhang, and D. Aron. 2022a. *The Concept and Empirical Evidence of SWIFT Methodology*. World Bank, Washington DC. https://elibrary.worldbank.org/doi/epdf/10.1596/38095.

Yoshida, N., S. Takamatsu, D. Aron, and K. Zhang. 2022b. "Poverty projections and profiling using a SWIFT-COVID-19 package during the COVID-19 pandemic." Mimeo. World Bank, Washington DC.

Appendix 1. SWIFT models for experiments

Table A1.1. SWIFT Model for Saint Lucia

	(1) ln(Consumption)
District==Castries City	-0.225***
	(0.0520)
District==Castries Sub-Urban	-0.155***
	(0.0461)
District==AnseLaRayeCanaries	-0.182*
	(0.0788)
Dependency Ratio	-0.313***
	(0.0614)
District==Vieuxfort	-0.315***
	(0.0621)
Urban/Rural	0.298***
	(0.0465)
Household Size	-0.216***
	(0.0195)
Household Size Sq.	0.840***
	(0.159)
Sex of Head (1=male)	0.105**
	(0.0324)
Age of Head	0.00467***
	(0.00117)
you went without eating for a whole day	-0.225**
	(0.0720)
your household ran out of food because of a lack of money	-0.378***
	(0.0448)
Employment Status	0.165***
	(0.0374)
Constant	9.614***
	(0.0986)
Observations	1474
Adjusted R^2	0.345

Standard errors in parentheses
$^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$
Source: authors' estimation using HBS (2016) in Saint Lucia.

Table A1.2 The SWIFT Model for HFPS Uganda Refugees

	(1) Log(Consumption)
REGION==West Nile	-0.826***
	(0.0962)
REGION==Southwest	-0.686***
	(0.0969)
Arrived before 2008	0.516***
	(0.108)
Arrived between 2008 and 2015	0.225**
	(0.0716)
Income Source 12M: Remittances	0.188**
	(0.0576)
Household Size	-0.130***
	(0.0336)
HH Size Squared	0.00667**
	(0.00232)
Scores for component 1	-0.316***
	(0.0786)
Constant	12.18***
	(0.142)
Observations	798
Adjusted R^2	0.323

Standard errors in parentheses
$^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$
Source: authors' estimation using URHS (2018) for Uganda Refugees.

Table A1.3 SWIFT Model for Rural Rwanda HFPS Round 2

	(1) ln(Consumption)
region==Western rural	-0.120***
	(0.0143)
region==Northern rural	-0.0678***
	(0.0150)
Household Size	-0.224***
	(0.0147)
Household Size Sq.	1.262***
	(0.138)
Head Sex	-0.216***
	(0.0201)
head age (square)	0.00457***
	(0.000413)
head_mrstat==Widow or widower	0.103***

	(0.0236)
Worked at least 1h in last 7 days	0.323***
	(0.0241)
head_sec==2 Private farm	-0.211***
	(0.0156)
consume local/imported rice in the past 7 days	0.236***
	(0.0127)
Constant	12.62***
	(0.0430)
Observations	12054
Adjusted R^2	0.170

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Source: authors' estimation using EICV5 (2017) in rural Rwanda.

Table A1.4 SWIFT Model for Urban Rwanda HFPS Round 2

	(1)
	ln(Consumption)
region==Kigali	0.347***
	(0.0450)
Household Size	-0.168***
	(0.0445)
Household Size Sq.	0.828*
	(0.379)
Head Sex	-0.110*
	(0.0486)
Head Age	0.00682***
	(0.00167)
Worked at least 1h in last 7 days	0.275***
	(0.0752)
head_sec==2 Private farm	-0.441***
	(0.0599)
consume local/imported rice in the past 7 days	-0.158***
	(0.0415)
consume beef meat in the past 7 days	0.701***
	(0.0489)
Constant	12.76***
	(0.133)
Observations	2526
Adjusted R^2	0.304

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.5 SWIFT Model for Ethiopia (Rural)

	(1) ln of per adult equivalent consumption
Region==TIGRAY	-0.333***
	(0.0466)
Region==AMHARA	-0.355***
	(0.0273)
Region==SNNP	-0.242***
	(0.0275)
Reduced Consumption Preferred Foods Past 7 Days	-0.171***
	(0.0303)
Reduce number of meals eaten in a day	-0.0240
	(0.0345)
Anyone in HHs is in agriculture last week	0.116**
	(0.0387)
Anyone in HHs is in non-farm business last week	0.247***
	(0.0401)
Household Size	-0.211***
	(0.0199)
Household size squared	0.0808***
	(0.0150)
Number of rooms(excluding kitchen, toilet and bath room)	0.0737***
	(0.0108)
Floor==cement creed	0.290***
	(0.0828)
Floor==tiles	0.310
	(0.161)
Floor==other	0.227**
	(0.0733)
Constant	10.02***
	(0.0672)
Observations	3115
Adjusted R2	0.195

Standard errors in parentheses
$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$
Source: authors' estimation using ESS4 in rural Ethiopia.

Table A1.6 SWIFT Model for Ethiopia (Urban)

	(1) ln of per adult equivalent consumption
Region==SNNP	-0.119***

	(0.0245)
Reduced Consumption Preferred Foods Past 7 Days	-0.0378
	(0.0305)
Reduce number of meals eaten in a day	-0.263***
	(0.0334)
Anyone in HHs is in agriculture last week	-0.119***
	(0.0250)
Anyone in HHs is in non-farm business last week	0.0866***
	(0.0199)
Household Size	-0.241***
	(0.0139)
Household size squared	0.101***
	(0.0113)
Head can read and write	0.233***
	(0.0214)
Number of rooms(excluding kitchen, toilet and bath room)	0.0995***
	(0.00794)
Floor==cement creed	0.277***
	(0.0206)
Floor==tiles	0.461***
	(0.0377)
Floor==other	0.228*
	(0.0942)
Constant	10.19***
	(0.0425)
Observations	3655
Adjusted R^2	0.395

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Source: authors' estimation using ESS4 in urban Ethiopia.

Appendix 2. Propensity Score Regressions in Reweighting

Table A2.1 Propensity Score Weighting for the St Lucia experiment

	Round 2 – PSW
Urban/Rural	0.0676
	(0.0886)
Household Size	0.120*
	(0.0604)
Dependency Ratio	-0.338*
	(0.150)
Household Size Sq.	-1.103
	(0.637)

Gender of the Head	-0.0588
	(0.141)
Internet Access	0.145
	(0.0907)
Constant	-0.388*
	(0.167)
N	2702

Note: The propensity score weighting procedure for the second round includes six variables: **household size, locality(urban/rural), dependency ratio, household size squared, sex of household head, and internet access of the household**.

$^* p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$

Source: authors' calculation using HBS 2016 in Saint Lucia and phone owners of the same survey.

Table A2.2. PSW Regressions for the Rwanda experiment

	Round 2
Household Size	0.240***
	(0.0253)
Dependency Ratio	-0.846***
	(0.0634)
Household Size Sq,	-1.347***
	(0.232)
Urban Share	0.0828
	(0.0447)
Age of Household Head	-0.00268**
	(0.000951)
District 1	0.0471
	(0.0573)
District 2	-0.0491
	(0.0397)
District 3	0.0182
	(0.0419)
District 4	0.0364
	(0.0443)
wallC1	0.119***
	(0.0303)
lightingC1	0.342*
	(0.143)
lightingC2	-0.387*
	(0.156)
TV	0.0150
	(0.0738)
Bicycle	0.307***
	(0.0400)
Decoder	-0.120
	(0.0830)

Cupboard	0.0777
	(0.0510)
N	24163
R squared	0.0249

Table A2.3. PSW regressions for the Uganda refugee experiment

	National PSM
Household size	0.208*
	(0.102)
Dependency ratio	-1.078**
	(0.375)
Household size squared	-0.0104
	(0.00673)
District = Kampala	0.361
	(0.339)
District = West Nile	-0.153
	(0.155)
Furniture ownership	0.332*
	(0.149)
Motorcycle ownership	0.452
	(0.537)
Bicycle ownership	0.156
	(0.276)

Number of rooms	0.0641
	(0.0901)
Constant	-1.081***
	(0.286)
N	1243
Pseudo R-sq.	0.0290

Table A2.4. PSW regressions for the Ethiopia experiment

	National PSW
Head age	0.0198*
	(0.00940)
Head age sq.	-0.000153
	(0.0000959)
Head sex	0.0278
	(0.0646)
Head marital status = married	-0.0312
	(0.0674)
Head education: incomplete primary school	0.349***
	(0.0709)
Head education: complete primary school	0.389***
	(0.109)
Head education: incomplete secondary school	0.455***
	(0.0921)
Head education: complete secondary school	0.497***
	(0.112)
Head education: post-secondary school	0.526***

	(0.0857)
Head labor market status: unemployed	-0.0581
	(0.0973)
Head labor market status: inactive	-0.128
	(0.0656)
Head engaged in agriculture	-0.185*
	(0.0808)
Head engaged in casual labor activities	0.0226
	(0.124)
Maximum remittances	0.0334
	(0.0532)
House ownership	-0.107
	(0.0599)
Number of people per room	0.0607
	(0.0408)
Urban/rural status	0.362***
	(0.0713)
Consumption quintile = 2	0.297**
	(0.0996)
Consumption quintile = 3	0.445***
	(0.0968)
Consumption quintile = 4	0.414***
	(0.0969)
Consumption quintile = 5	0.481***
	(0.0980)
Household size	-0.00814
	(0.0141)
Constant	-2.240***
	(0.231)
N	9301

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Source: authors' calculation using ESS4 and the subsample phone owners of the same phone survey who were also reached in HFPS round 7 in Ethiopia.

Appendix 3. Sampling weight adjustments when including the population shares by country of origin in the matching conditions of maxentropy

It is straightforward to match household level averages between a phone survey and a reference survey using maxentopy, but it is not as simple to add individual level averages like population shares as matching conditions. Maxentropy can match household level averages or individual level averages one at a time; when all matching conditions are defined as household level averages (like average household size), the maxentropy command adjusts household weights in a way that makes target indicators equate the statistics in the reference survey. Under this circumstance, if we include matching conditions defined with population averages, an additional adjustment is required for the population-adjusted variables before these variables can be added into this maxentropy command which was initially aimed to adjust household weights. This appendix proves why such an adjustment is needed.

If we include household-level averages of K variables in the matching conditions, the STATA command of maxentropy searches for a set of household weights $\{w_i\}_{i=1}^N$, where N is the sample size of the phone survey that maximizes the objective function while satisfying the following conditions:

$$\frac{\sum_i x_i^{kp} * w_i}{\sum_i w_i} = \bar{x}^{kr} \qquad for\ k = 1, \dots, K \qquad \text{Eq. 1}$$

where x_i^{kp} is a value of household i's kth variable for household (r) in the phone survey (p) and $\bar{x}^{kr}$ a target value for the kth variable estimated in the reference survey(r).

We now aim to match a population share of refugees who came from country j to the counterpart of the reference survey $(\bar{s}_j^r)$, which is defined as:

$$\bar{s}_j^r = \frac{\sum_i CO_i^{jr} * w_i^r * hhsize_i^r}{\sum_i w_i^r * hhsize_i^r} \qquad \text{Eq. 2}$$

where w_i^r is the household weight for household i, $hhsize_i^r$ the household size of household i, and CO_i^{jr} a dummy that takes 1 if household i comes from the origin country j and 0 otherwise. All these numbers are drawn from the reference survey (r).

If we include an indicator for the country of origin, j, in the phone survey, CO_i^{jp}, into the maxentropy command (Eq. 1), the matching condition then is defined as:

$$\frac{\sum_i CO_i^{jp} * w_i}{\sum_i w_i} = \bar{s}_j^r \qquad \text{Eq. 3}$$

where CO_i^{jp} refers to an indicator that a refugee in the phone survey came from country j. Then, the left-hand-side expression of Eq. 3 represents the share of **households** with the country of origin being j, while the target value in the reference survey on the right-hand side is the population share with country j being their origin. To avoid such a mismatch between the reference and phone surveys, instead of replacing x_i^k in Eq. 1 with CO_i^k as is the case in Eq. 3, we insert $CO_i^{jp} * hhsize_i^r$ into x_i^k, and the resulting expression can be seen in Eq. 4:

$$\frac{\sum_i CO_i^{jp} * hhsize_i^p * w_i}{\sum_i w_i} = \bar{s}_j \qquad \text{Eq. 4}$$

However, comparing Eq. 2 and Eq. 4, we observe that the left-hand side of Eq. 4 is not exactly a population average, whose denominator should be the total number of individuals as is the case in Eq. 2. Therefore, to make the left-hand side expression of Eq. 4 a population share defined by country of origin, j, as in Eq. 2, we insert the following variable into Eq. 1, and derive Eq. 5 as below:

$$CO_i^{jp} * hhsize_i^p * \frac{\sum_i w_i^0}{\sum_i w_i^0 \times hhsize_i^p} \qquad (A2-1) \qquad \text{Eq. 5}$$

where $\{w_i^0\}$ is a set of initial household weights. Regarding the choice of initial weights, if there are pre-existing household weights in the phone survey, these are used as the initial household weights. If, on the other hand, there is no initial household weight in the phone survey, then we create initial weights by executing the maxentropy procedure in the phone survey with all household-level indicators being the targets and without including any of the population-level averages. The generated weights are then used as the initial household weights.

Combining Eq. 1 and Eq. 5, we observe that the matching condition of this variable becomes:

$$\frac{\sum_i CO_i^{jp} * hhsize_i^p * \frac{\sum_i w_i^0}{\sum_i w_i^0 \times hhsize_i^p} * w_i}{\sum_i w_i^*} = \bar{s}_j \qquad (A2-2) \qquad \text{Eq. 6}$$

Maxentropy searches for the optimal solution $\{w_i^*\}$ that maximizes entropy conditional on all matching conditions, including the matching of all household-level averages and population-level averages. If $\{w_i^*\} = \{w_i^0\}$, then Eq. 6, the condition for matching the share of the population with the country of origin being j, can be rewritten as Eq. 7. :

$$\frac{\sum_i CO_i^{jp} * w_i^* * hhsize_i^p}{\sum_i w_i^* * hhsize_i^p} = \bar{s}_j \qquad (A2-3) \qquad \text{Eq. 7}$$

However, $\{w_i^*\} = \{w_i^0\}$ is not necessarily satisfied. We do not know the optimal weights $\{w_i^*\}$, and would have to start with other weights, such as $\{w_i^0\}$ in the analysis. If $\{w_i^*\} \neq \{w_i^0\}$, then Eq. 6 cannot be simplified to be Eq. 7, and the left-hand-side expression of Eq. 6 would inevitably deviate from the population share as is defined by Eq. 7. To minimize the bias, it is important to choose $\{w_i^0\}$ to be as close to $\{w_i^*\}$ as possible.

In practice, it is not possible to target $\{w_i^*\}$ immediately; as a remedy, the World Bank (2021) first calculates the initial weights $\{w_i^0\}$ by conducting the maxentropy exercise only, including household-level indicators and excluding the population shares by country of origin. Applying these weights as the initial weights, they then construct a variable defined in Eq.5 and conduct another round of maxentropy where the population shares by country of origin are included. The final weights they obtain are $\{w_i^*\}$, which matches all included conditions

at each region exactly. This suggests the weights derived by the World Bank (2021) are indeed the target weights, $\{w_i^*\}$, which can match population-level averages to be identical between the reweighted phone survey and the reference survey.

Appendix 4. Methods and Formula for Maximum Entropy Reweighting in Stata (Wittenberg 2010, section 3.5)

Instead of solving the constrained optimization problem given by the first-order conditions [(3) to (5)] or their cross-entropy analogs, (Golan, Judge, and Miller 1996) show that the solution can be found by maximizing the unconstrained dual cross-entropy objective function.

$$L(\lambda) = \sum_{j=1}^{J} \lambda_j y_j - \ln\{\boldsymbol{\Omega}(\lambda)\} = M(\lambda) \quad (8)$$

where $\boldsymbol{\Omega}(\lambda)$ is given by (7). Golan, Judge, and Miller show that this function behaves like a maximum likelihood. In this case,

$$\nabla_\lambda \boldsymbol{M}(\lambda) = \boldsymbol{y} - \boldsymbol{X}'\boldsymbol{p} \quad (9)$$

so that the constraint is met at the point where the gradient is zero. Furthermore,

$$-\frac{\partial^2 \boldsymbol{M}}{\partial \lambda_j^2} = \sum_{i=1}^{n} p_i x_{ji}^2 - \left(\sum_{i=1}^{n} p_i x_{ji}\right)^2 = var(x_j) \quad (10)$$

$$-\frac{\partial \boldsymbol{M}}{\partial \lambda_j \partial \lambda_k} = \sum_{i=1}^{n} p_i x_{ji} x_{ki} - \left(\sum_{i=1}^{n} p_i x_{ji}\right)\left(\sum_{i=1}^{n} p_i x_{ki}\right) = cov(x_j, x_k) \quad (11)$$

where the variances and covariances are taken with respect to the distribution **p.** The negative of the Hessian of **M** is therefore guaranteed to be positive definite, which guarantees a unique solution provided that the constraints are not inconsistent.

Golan, Judge, and Miller (1996) note that the function **M** can be considered an expected log-likelihood, given the exponential family **p(λ)** parameterized by λ. Wittenberg (2010) uses Stata's maximum likelihood routines to estimate λ, giving it the dual objective function [(8)], gradient [(9)], and negative Hessian [(10) and (11)]. Because of the globally concave nature of the objective function, convergence should be relatively quick, provided that there is a feasible solution in the interior of the parameter space. The Stata command that calculates these is contained in maxentlambda_d2.ado. The command checks for some obvious errors. For example, the population means (y_j) must be inside the range of the X_j variables. If any mean is on the boundary of the range, then a degenerate solution is feasible, but the corresponding Lagrange multiplier will be $\pm\infty$, so the algorithm will not converge.

Appendix 5. The performance of the SWIFT methodology in projecting poverty rates

Yoshida et al. (2022) show the SWIFT methodology is empirically proven to be reliable using multiple "out-of-sample" tests. The SWIFT models are trained using a household survey dataset called "training data." If the performance of the SWIFT model is evaluated using the training data, the test is called a "within-sample" test. The within-sample test can be biased due to over-fitting and model instability. The over-fitting issue means that even though a model estimates the poverty rate of the training data extremely well, the model might not be able to estimate poverty rates outside the training data. The stability issue can result in a model producing a large bias in poverty estimates if used in datasets significantly older or newer than the training data because the model parameters change between the training data and other datasets.

Table A5.1. Comparisons of poverty rates between actual consumption and SWIFT imputed expenditures (%)

Countries & years	Actual consumption		SWIFT Imputation (second year)
	First-year	Second year	
Vietnam (1992/93 - 97/98)*	60.6	37.4	36.7
Inner Mongolia (2000 - 2004)*	19.0	6.2	7.8
Kenya (1997 - 2005/6)*	50.8	46.6	45.5
Russian Federation (1994 – 2003)*	11.4	11.1	9.2
Morocco (2001-2007)**	15.3	8.9	8.4
Afghanistan (2011 – 2016)	38.3	54.5	53.5
Albania (2005-2008)	17.7	12.1	13.0
Malawi (2005-2011)	51.6	50.2	49.7
Romania (2011-2012)	22.6	21.7	22.3
Rwanda (2005-2008)	56.7	44.9	43.3
Sri Lanka (2009-2012)	8.7	6.5	7.0
Uganda (2009-2012)	24.5	19.5	23.3

Source. * Christiaensen et al. (2012); ** Douidich et al. (2013); all others from Yoshida et al. (2022).

The SWIFT approach uses cross-validation tests to ensure all SWIFT models do not have the overfitting issue. To see whether the SWIFT approach is vulnerable to the model stability issue, Yoshida et al. (2022) used Christiaensen et al. (2012) test, where a model is trained in one round of the household survey, and the model performance is tested in another round of the household survey. If a model is not stable over time, the model performance should be bad. Christiaensen et al. (2012) test is an accurate test for the model stability, but its problem is that this test can be done only if we have two household surveys with comparable consumption/income and regression variables. Unfortunately, two subsequent household surveys are not always comparable in many developing countries because the government often changes the questionnaire, survey logistics, and training of enumerators when it launches a new survey. Yoshida et al. (2022) selected two subsequent household surveys that are confirmed to be comparable by the World Bank's country poverty economists and examined the performance (see Table A5.1).

Table A5.1 shows that the performance of SWIFT models is quite good. Except for Uganda (2009-2012), the differences between actual and imputed poverty rates are less than two percentage points in absolute terms. The first five results are drawn from Christiaensen et al. (2012) and Douidich et al. (2013), where models used household demographics, asset ownership, and housing conditions. However, Yoshida et al. (2022) found that

these variables respond to changes in the socioeconomic environment, and the poverty estimates from these models are not accurate when a country faces a big crisis just before the second round of household surveys is collected. This was the case for Afghanistan (2011 – 2016), where a big conflict occurred just before the 2016 survey was collected. As a result, if a model includes only the slow-changing variables listed above, the model severely underestimates the actual poverty rate of 2016. To mitigate this risk of model stability issue, Yoshida et al. (2022) proposed a new SWIFT approach called SWIFT Plus, where a model includes some fast-changing variables, like consumption dummies, subjective wellbeing, and food security variables. With the SWIFT Plus approach, the new model estimated the 2016 poverty rate within a margin of one percentage point. The last seven cases in Table 1 used the SWIFT Plus approach.