HHS Public Access

Author manuscript

J Surv Stat Methodol. Author manuscript; available in PMC 2024 July 30.

Published in final edited form as:

J Surv Stat Methodol. 2021 November ; 9(5): 1035–1049. doi:10.1093/jssam/smaa020.

DEALING WITH INACCURATE MEASURES OF SIZE IN TWO-STAGE PROBABILITY PROPORTIONAL TO SIZE SAMPLE DESIGNS: APPLICATIONS IN AFRICAN HOUSEHOLD SURVEYS

GRAHAM KALTON,
Westat, 1600 Research Blvd, Rockville, MD 20850, USA

ISMAEL FLORES CERVANTES[*],
Westat, 1600 Research Blvd, Rockville, MD 20850, USA

CARLOS ARIEIRA,
Westat, 1600 Research Blvd, Rockville, MD 20850, USA

MIKE KWANISAI,
Westat, 1600 Research Blvd, Rockville, MD 20850, USA

ELIZABETH RADIN,
ICAP, 722 West 168th Street, New York, NY 10032, USA

SUZUE SAITO,
ICAP, 722 West 168th Street, New York, NY 10032, USA

ANINDYA K. DE,
U.S. Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333, USA

STEPHEN MCCRACKEN,
U.S. Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333, USA

PAUL STUPP
U.S. Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333, USA

Abstract

The units at the early stages of multi-stage area samples are generally sampled with probabilities proportional to their estimated sizes (PPES). With such a design, an overall equal probability (EP) sample design would yield a constant number of final stage units from each final stage cluster if the measures of size used in the PPES selection at each sampling stage were directly proportional to the number of final stage units. However, there are often sizable relative differences between the measures of size used in the PPES selections and the number of final stage units. Two common approaches for dealing with these differences are: (1) to retain a self-weighting sample design, allowing the sample sizes to vary across the sampled primary sampling units (PSUs) and (2) to retain the fixed sample size in each PSU and to compensate for the unequal selection probabilities by weighting adjustments in the analyses. This article examines these alternative designs in the

[*]Address correspondence to Ismael Flores Cervantes, Westat, 1600 Research Blvd, Rockville, MD 20850, USA; ismaelflorescervantes@Westat.Com.

context of two-stage sampling in which PSUs are sampled with PPES at the first stage, and an equal probability sample of final stage units is selected from each sampled PSU at the second stage. Two-stage sample designs of this type are used for household surveys in many countries. The discussion is illustrated with data from the Population-based HIV Impact Assessment surveys that were conducted using this design in several African countries.

Keywords

Clustering effect; Design effect; Equal probability sample; Equal subsample size; Weighting effect

1. INTRODUCTION

In most multi-stage surveys, probability proportional to size (PPS) sampling is used to select the units up until the final stage of sampling. In practice, the measures of size used for the PPS sampling are generally only estimates of the actual number of elements in the units at the various stages, and a more accurate description of the method is, therefore, probability proportional to estimated size (PPES) sampling. With PPES sampling, a common choice to be made for the final stage of sampling is between sampling a fixed number of final stage units in each selected final stage cluster with weighting adjustments made to compensate for unequal selection probabilities or applying a sampling fraction that maintains equal weights for the elements. This article addresses this general choice in the context of a two-stage sample design, but the issues discussed apply more generally to PPES designs with more than two sampling stages. The discussion is illustrated using a set of household surveys conducted in Africa.

Two-stage sample designs are used for household surveys in many countries, with the primary sampling units (PSUs) being sampled with PPES, where the measures of size are estimates of quantities such as the numbers of households or persons in the PSU. Samples of households are selected, generally by systematic sampling, from the lists of households compiled or otherwise obtained for the sampled PSUs. In almost all cases, the measures of size are inaccurate measures of the PSUs' current sizes, generally because they are based on out-of-date census data. The inaccuracies can be substantial when the last census was carried out several years earlier, when some areas have undergone major recent development, or have experienced major disasters such as cyclones, tsunamis, or civic unrest. Issues of inaccurate measures of size also arise when the size measures relate to a variable that is only indirectly associated with the true sizes of the sampling units; for example, in the United States, the measure of size of a segment for a face-to-face interview survey may be the count of the households in the segment at the last census, whereas the count of interest is the segment's current number of residential addresses on the US postal lists.

With a two-stage design, if the current PSU sizes are known and a PPS sample is selected, the probability of household β in PSU α being included in the sample is

$$P(\alpha\beta) = \left(\frac{aN_\alpha}{\Sigma^A N_\alpha}\right)\left(\frac{b_\alpha}{N_\alpha}\right) = \left(\frac{ab_\alpha}{\Sigma^A N_\alpha}\right),$$

(1)

where a PSUs are selected from A PSUs at the first stage with probabilities proportional to sizes N_α; and b_α households are selected with equal selection probabilities at the second stage from the N_α households in that PSU. If $b_\alpha = b$; then the overall selection probability $P(\alpha\beta) = f$; a constant. This sample design is statistically efficient–for a given total sample size, the design effect from clustering is smallest when an overall epsem sample is selected by sampling PSUs with PPS and selecting the same subsample size in each sampled PSU (see later). It is also operationally attractive because it produces the same interviewing workload in each sampled PSU.

In practice, however, the PPS design is not feasible because the current sizes of the PSUs N_α are not known. Instead, the PSUs are sampled with probabilities proportional to estimated sizes M_α so that the overall household selection probabilities are given by equation (1) as

$$P(\alpha\beta) = \left(\frac{aM_\alpha}{\Sigma^A M_\alpha}\right)\left(\frac{b_\alpha}{N_\alpha}\right) = \frac{ab_\alpha}{MK_\alpha},$$

(2)

where b_α households are selected with equal probability (EP) at the second stage from the N_α households listed for sampled PSU α, $M = \Sigma^A M_\alpha$; and $K_\alpha = N_\alpha / M_\alpha$ is the ratio of the listed size of PSU α to its estimated size used in the PPES selection. If $K_\alpha = K$, the design is a PPS design, and with $b_\alpha = b$; the overall selection probability is $P(\alpha\beta) = f$; a constant. In general, however, K_α varies across the sampled PSUs.

As has been noted earlier, there are two common approaches for dealing with the problem of inaccurate measures of size.

- One approach selects a fixed sample size (FSS) of households (e.g., $b = 25$ households) from each selected PSU in each stratum (the fixed take may vary between strata). With the FSS design, households are sampled with unequal overall selection probabilities, thus requiring weighting compensation in the analyses.
- The second approach, termed the EP design, allows the sample sizes to vary across the sampled PSUs in a way that produces an overall EP (*epsem*) design.

With the FSS design, obtaining a specified sample size $n = ab$ is simply achieved by determining the number of PSUs to sample. However, achieving a specified sample size with the EP design faces two complexities. First, the overall sampling rate has to be chosen: an overall sampling rate of $P(\alpha\beta) = n/N$ (with $N = \Sigma^A N_\alpha$) is required to achieve a sample size of n, but N is generally unknown. Obtaining a good estimate of N may not be feasible, and differences between that estimate and the true N will result in a sample size that differs from that planned. Second, even if N were known exactly, the actual sample size still depends on the PSUs selected for the sample; the sample size in PSU α is $bN_\alpha / M_\alpha = bK_\alpha$, and the

overall sample size is $n_c = b\Sigma^a K_\alpha = ab\bar{K}_c$, where the subscript c denotes that the sample size is conditional on the set of PSUs in the sample.

This article describes a modification in the EP design that provides full control of the overall sample size, control that is lacking with the EP design. This modification starts with the selection of a PPES sample of PSUs in the standard way. An estimate of the population size is made from that first phase sample as

$$N_c = \sum_\alpha^a N_\alpha / P(\alpha) = \sum_\alpha^a M N_\alpha / aM_\alpha = M\bar{K}_c,$$

The required overall sampling fraction is then given by $P(\alpha\beta) = n/N_c = n/M\bar{K}_c$. Substituting this overall sampling fraction into (2) gives the sample size to be selected in PSU α as $b_\alpha = nMN_\alpha/aM\bar{K}_cM_\alpha$, whence $\sum_\alpha^a b_\alpha = n$. Since the estimated population size in the first phase sample N_c depends on the set of sampled PSUs, this two-stage two-phase sample design does not yield an epsem sample, and indeed the overall selection probabilities cannot be computed. However, it is a self-weighting (SW) sample design that, with equal weights for all sampled elements, produces unbiased estimates of population totals. See the discussion of the π^* estimator in Särndal, Swensson, and Wretman (1992, pp. 347–50) and Fuller (2009, pp. 215–7). Henceforth, this design is described as an SW design. The SW design is a close approximation to the EP design when $\bar{K}_c$ is almost constant across possible first stage samples of PSUs, as when a large number of PSUs is selected.

The article is organized as follows. Section 2 presents the theory for the comparison of the statistical efficiencies of the FSS and SW approaches. As the theory demonstrates, the relative efficiency depends primarily on the variability in the ratios of the current sizes to the measures of size used in the PPES selection (the K_αs) and, to a lesser extent, on the homogeneity of the variable y under study for a particular analysis within the PSUs, measured by the intraclass correlation ρ_y. Section 3 describes the SW sample design used in the Population-based HIV Impact Assessment (PHIA) surveys for which the relative efficiencies of the FSS and SW designs have been computed. Section 4 reports results for the components affecting the relative precision of the FSS and SW sample designs and provides estimates of this relative precision for these surveys. Section 5 presents some concluding remarks on the advantages and disadvantages of these two sample designs.

2. THEORETICAL RESULTS

With the FSS design for selecting a fixed number of households within sampled PSUs, and with epsem sampling of households within PSUs, a household's overall selection probability is given by (2) with $b_\alpha = b$. With this unequal probability sample design for sampling households, weights proportional to K_α are needed in the analysis. As shown later, the variability in these weights decreases the precision of the survey estimates.

The SW sample design avoids the variability in the weights caused by the variation in the K_α (still, in practice, the analysis weights will vary because of any departures from the ideal SW design and because of nonresponse and calibration adjustments). However, that

benefit comes at the price of variability in the numbers of households b_α selected in sampled PSUs, with b_α being proportional to K_α. This variability in b_α across the sampled PSUs presents fieldwork challenges in general and particularly with two-stage sample designs with geographically dispersed samples of PSUs. It is not cost-effective for traveling interviewers to conduct interviews and call-backs to unavailable or "soft" refusal households in PSUs with small samples of households. PSUs that have large sample sizes can also present logistical challenges in fieldwork management.

We assess the relative precision of survey estimates obtained under the SW and FSS designs by comparing the design effects of estimates of population means for the two designs. Two factors contribute to these design effects: clustering and weighting (Kish 1995). With equal-sized PSUs (clusters) and an overall epsem design, the clustering design effect for a sample mean is $Deff_c = 1 + (b - 1)\rho_y$, where b is the sample size in each PSU and ρ_y is the within PSU intraclass correlation for the outcome variable y. With an overall epsem design and variable within PSU sample size b_α, b is often replaced by $\bar{b} = \Sigma^a b_\alpha / a$ (see, e.g., Kish 1965). However, Holt (1980) and Skinner (1986) show that b is better replaced by $\Sigma^a b_\alpha^2 / \Sigma^a b_\alpha = \bar{b}\left[1 - cv^2(b)\right]$, where $cv(b)$ denotes the coefficient of variation of the b_α in the sample. (Note that throughout this article, $cv^2(x) = \text{var}(x)/\bar{x}^2$ with $\text{var}(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$ for a sample of size n, with a divisor of n rather than $n - 1$.)

Under the assumption that the weights are uncorrelated with the survey variables, the weighting design effect is $Deff_w = 1 + CV^2(w)$, where $CV(w)$ is the coefficient of variation of the sampling weights (Kish 1992). This lack of correlation assumption does not always hold, but it is widely accepted as a reasonable approximation for planning sample designs in situations like the current one. See Valliant, Dever, and Kreuter (2013, pp. 375–80) and Spencer (2000) for more discussion of the weighting design effect and the consequences of weights that are linearly correlated with a continuous survey variable.

Kish (1987) proposed modeling the overall design effect for a sample mean with an unstratified two-stage design and variable weights as the product of $Deff_c$ and $Deff_w$, that is, by

$$Deff = \left\{1 + CV^2(w)\right\} \times \left\{1 + (\bar{b} - 1)\rho_y\right\}. \tag{3}$$

In a model-based justification of Kish's formula, Gabler, Haeder, and Lahiri (1999) developed a modification in (3) that allows for unequal selection probabilities of elements within PSUs, replacing $\bar{b}$ by

$$b^* = \frac{\sum_{\alpha=1}^{a}\left(\sum_{\beta=1}^{b_\alpha} w_{\alpha\beta}\right)^2}{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b_\alpha} w_{\alpha\beta}^2},$$

where $w_{\alpha\beta}$ is the weight for household β in PSU α. With an EP sample of households within each sampled PSU, as assumed throughout this article, $w_{\alpha\beta} = w_\alpha$, $w_\alpha = MK_\alpha/ab_\alpha$

$$b^* = \frac{\sum_{\alpha=1}^{a}(b_\alpha w_\alpha)^2}{\sum_{\alpha=1}^{a} b_\alpha w_\alpha^2},$$

and

$$deff_w = 1 + cv^2(w) = 1 + \frac{\mathrm{var}(w)}{\overline{w}^2} = \frac{n\Sigma^a b_\alpha w_\alpha^2}{\left(\Sigma^a b_\alpha w_\alpha\right)^2},$$

where $deff_w$ is an estimate of $Deff_w$, $n = \Sigma^a b_\alpha$, $\overline{w} = \Sigma^a b_\alpha w_\alpha/n$, and $\mathrm{var}(w) = \Sigma^a b_\alpha(w_\alpha - \overline{w})^2/n$. If the households are selected with an SW design, $b^* = \Sigma^a b_\alpha^2/\Sigma^a b_\alpha = \bar{b}[1 + cv^2(b)]$, in agreement with the result of Holt and Skinner. Note that for a given $\bar{b}$, $Deff_c$ is smallest when $cv(b) = 0$, as occurs with exact PPS sampling.

Consider any sample allocation of the b_α elements at the second stage subject to a fixed overall sample size of $n = \Sigma^a b_\alpha$. Then, $w_\alpha \propto K_\alpha/b_\alpha$. Hence, the estimates of $Deff_w$ and $Deff_c$ for the same set of sampled PSUs are given by

$$deff_w = 1 + cv^2(w) = n\left(\Sigma^a K_\alpha^2/b_\alpha\right)/\left(\Sigma^\alpha K_\alpha\right)^2,$$

$$deff_c = \left\{1 + \left(b^* - 1\right)\hat{\rho}_y\right\} = \left\{1 + \left[\Sigma^a K_\alpha^2/\Sigma^a\left(K_\alpha^2/b_\alpha\right) - 1\right]\hat{\rho}_y\right\},$$

and the overall estimated design effect is

$$deff = \frac{n(1-\hat{\rho}_y)\Sigma^a\left(K_\alpha^2/b_\alpha\right)}{\left(\Sigma^a K_\alpha\right)^2} + n\hat{\rho}_y\frac{\Sigma^a K_\alpha^2}{\left(\Sigma^a K_\alpha\right)^2}.$$

For a given total sample size n, the estimated overall design effect $deff$ is minimized when $\Sigma^a K_\alpha^2/b_\alpha$ is minimized. The Cauchy inequality states that $\left(\Sigma_h a_h^2\right)\left(\Sigma_h b_h^2\right) \geq \left(\Sigma_h a_h b_h\right)^2$, with the minimum occurring when $a_h/b_h = C$, a constant. Applying this inequality to $\left(\Sigma^a K_\alpha^2/b_\alpha\right)\left(\Sigma^a b_\alpha\right)$ gives a minimum value for $deff$ when $\left(K_\alpha b_\alpha^{-1/2}\right)/b_\alpha^{1/2} = C$, that is, $b_\alpha \propto K_\alpha$. Thus, the SW design with $b_\alpha \propto K_\alpha$ minimizes $deff$ and, hence, it is a more efficient design than the FSS design for any set of sampled PSUs.

We turn now to the relative efficiency of the FSS and SW designs. With the FSS design, $b^* = b$ and, with $w_{\alpha\beta} \propto K_\alpha$, $1 + cv^2(w) = 1 + cv^2(K)$. With the SW design, $b_\alpha = bK_\alpha/\overline{K}$, and $w_{\alpha\beta} = w_\alpha$, then $\left\{1 + cv^2(w)\right\} = 1$, and b^* reduces to $\Sigma^a b_\alpha^2/\Sigma^a b_\alpha$. Hence

$$b^* = \left[b^2\Sigma^a K_a^2/\bar{K}_c^2\right]/\left[b\Sigma^a K_a/\bar{K}_c\right] = b\Sigma^a K_a^2/a\bar{K}_c^2 = b\left[1+cv^2(K)\right].$$

Note that the term $\left\{1+cv^2(K)\right\}$ in the SW design is equal in magnitude to $F=\left\{1+cv^2(w)\right\}$ that applies to the FSS design.

In summary, for an unstratified two-stage sample design,

$$deff_{fss} = F\{1+(B-1)\hat{\rho}_y\}$$

and

$$deff_{sw} = \{1+(Fb-1)\hat{p}_y\}.$$

Hence,

$$deff_{fss} - deff_{ep} = (F-1)(1-\hat{\rho}_y) \geq 0,$$

which implies that the ratio of design effects can be expressed as

$$\hat{R} = \frac{deff_{fss}}{deff_{sw}} = 1 + \frac{(F-1)(1-\hat{\rho}_y)}{1+(Fb-1)\hat{\rho}_y} = 1 + \frac{cv^2(K)(1-\hat{\rho}_y)}{1+(b^*-1)\hat{\rho}_y}. \quad (4)$$

The quantity $\hat{R}$ estimates the ratio of the variance estimates of the sample means for the two designs with the same sample size and the same sample of PSUs. Thus, for the sample means to attain the same precision, the sample size for the FSS design needs to be larger than that for the SW design by this ratio. In particular, this ratio is relatively large when the variability in the K_α is large, $\hat{\rho}_y$ is small (say, 0.05 or even 0.10), and b is small.

3. THE SAMPLE DESIGNS FOR THE PHIA SURVEYS

The sample designs of each of the thirteen PHIA surveys covered in this article focused on two main objectives: to estimate with specified levels of precision (a) the incidence of HIV nationally and (b) the subnational proportions of HIV persons whose HIV viral load was suppressed as a result of antiretroviral therapy (ART). To satisfy the second objective, the overall sample was stratified into major domains by relevant subdivisions of the country (such as region or province), and the total sample size was allocated to these domains in a manner designed to achieve the specified precision goals. Such an allocation often resulted in the use of overall household sampling rates that varied from domain to domain. A two-stage sample design was implemented in each country, with PSUs selected with probabilities proportional to the numbers of households in the PSU according to the previous population census. Lists of households were compiled for each of the sampled PSUs, and

systematic samples of households were selected from the lists. All adults in a specified age range in the selected households were included in the study sample. Children were subsampled, and in some countries, adults were subsampled for additional data collection modules for special studies. This article deals only with the adult population of primary interest. The age range of interest starts at 15 years of age in all countries, but the upper age limit varied across countries. The analyses reported here relate to persons between the ages of fifteen and forty-nine years, an age range covered in all the surveys.

The PSUs were generally the enumeration areas (EAs) delineated for the last census. However, some EAs were too large to be listed entirely. Sometimes a PSU was large at the time of the previous census (and remained too large for listing entirely) and sometimes the PSU was found to be a growth area that had become too large in the time since the census; in the first case the PSU's size would have been reflected in its PPES selection probability, but that would not be so in the second case. In both cases, an intermediate stage of sampling known as segmentation was introduced. The large sampled PSUs were divided into segments, one segment was selected in each PSU, and the listing operation was carried out only in the sampled segment. Let $\widehat{M}_{\alpha\beta} = p_{\alpha\beta}M_{\alpha}$ be the estimated measure of size of segment β in PSU α, where $p_{\alpha\beta}$ is an estimate of the proportion of households in that segment in that PSU computed from an initial scouting operation carried out to make a rough count of the numbers of dwellings in each segment (with $\Sigma_{\beta}p_{\alpha\beta} = 1$ and hence $\Sigma_{\beta}\widehat{M}_{\alpha\beta} = M_{\alpha}$). Then, the conditional selection probability for segment $\alpha\beta$ is $\widehat{M}_{\alpha\beta}/M_{\alpha}$, and its overall selection probability is $a\widehat{M}_{\alpha\beta}/M$. Since a single segment is sampled from each segmented PSU, and PSUs are sampled by systematic PPES sampling, the design is equivalent to a two-stage sample in which the segments are the PSUs that are sampled with probabilities proportional to their estimated census counts $\widehat{M}_{\alpha\beta}$.

4. RELATIVE PRECISION $\widehat{R}$ BASED ON SEVERAL PHIA SURVEYS

It can be seen from (4) that the magnitude of the ratio of the design effects $\widehat{R}$ depends on the values of the estimates $cv^2(K)$, $\hat{\rho}_y$, and the average sample size per PSU b^*. The next two subsections report the estimates $cv^2(K)$ and $\hat{\rho}_y$ for a number of PHIA surveys, and the last subsection then examines the values of $\widehat{R}$ for a variety of survey estimates of population percentages in the PHIA surveys.

The assumption underlying the weighting design effect $\left\{1 + cv^2(w)\right\}$ in (4) is that the weights–which are proportional to the K_{α}s in the FSS design–are not correlated with the survey variables. We performed some calculations in three PHIA surveys to evaluate this assumption for four key HIV-related variables that influence the sample design: testing positive for HIV, ever having been test for HIV, on ART for those testing positive for HIV, and with viral load suppression for those on ART. Correlations between each of these variables and K_{α} were uniformly small, all falling within the range of −0.04 and +0.04. Based on these findings, we conclude that $\left\{1 + cv^2(w)\right\}$ serves as a reasonable estimate of the weighting design effect.

A large number of PSUs have been selected in the PHIA surveys, ranging from around 400 to over 500 PSUs in most surveys. As a result, the sample $\bar{K}_c$ values have little variance around the population mean of N/M. In such a situation, the SW design is approximately an epsem design, corresponding to the EP design. In this case, the benefit of the SW design over the EP design is that it avoids the need to obtain an external estimate of the population size $N = \Sigma^A N_a$, as required for determining the overall sampling fraction $f = n/N$ for the EP design. Moreover, by sampling households from the actual listings, the SW design automatically makes an allowance for the noncoverage that occurs with the household listing operation. This allowance for noncoverage has to be incorporated in the calculation of the overall sampling fraction with the EP design.

4.1 Values of $\bar{K}_c$ and $cv^2(K)$

As crude indicators of the growth in the number of households in a country since the last census, the third column of table 1 presents national averages of the within-domain K_as, $\bar{K}c = \Sigma^a K_a/a$. These national averages were calculated separately for the regional (or provincial) domains of each of the thirteen countries and then averaged across the domains, with the averaging being done in a way that retained the domain stratification but removed the PHIA disproportionate allocation across domains. It should be noted that these averages are imperfect measures of growth because they reflect a combination of factors such as: the growth in the number of households since the last census; listed units that were unoccupied at the time of survey data collection; listed units that contained more than one household; and some degree of undercoverage in the listings. As might be expected, the magnitude of the $\bar{K}_c$s varies by domain; for example, the regional $\bar{K}_{ch}$s range from 1.25 to 2.37 for country B, from 1.31 to 1.56 for country F, and from 0.98 to 1.19 for country L.

As with the calculations of $\bar{K}_c$, the estimates $cv^2(K)$ presented in the last column of table 1 are national average within-domain values, with compensation for the disproportionate allocation across domains; thus, these estimates $cv^2(K)$ are applicable for a design with a proportionate allocation with the domains as strata. With the FSS sample design, the estimates $cv^2(K) = cv^2(w)$ are important for planning purposes because they are needed for determining the overall sample size required to meet the specified precision levels for the survey estimates. The magnitude of $cv^2(K)$ has only a minor effect on the sample size needed for the SW design (through b^*).

The countries are listed in table 1 in order of the magnitude of $cv^2(K)$. The average $cv^2(K)$ s show considerable variability across countries. There is no clear-cut relationship between the values of the $cv^2(K)$ s and the recency of the last census, although, as may be expected, the $cv^2(K)$ are in the lower part of the range for countries I, J, and K that had censuses within three years of the survey.

The PHIA surveys are designed to produce HIV-related measures at specified levels of precision for regional domains as well as for the nation. The regional values of $cv^2(K)$ are, therefore, important. These values vary markedly across regions. For example, the regional

$cv^2(K)$ s range from 0.07 to 0.48 for country B, from 0.13 to 0.33 for country F, and from 0.03 to 0.15 for country L.

4.2 Values of the Estimates of the Intraclass Correlations $\widehat{\boldsymbol{\rho}}_{\mathbf{y}}$

As can be seen in (4), the value of $\hat{R}$ depends on the estimates of the intraclass correlation coefficient $\hat{\rho}_y$ for the particular variable under study. To examine the magnitude of the intraclass correlations, stratum (domain)-level estimates of $\hat{\rho}_y$ for stratum h were computed using the approach in a paper by Chen and Rust (2017) that extends the Gabler, Haeder, and Lahiri (1999) design effect model given by (4) to two- and three-stage designs with stratification. Chen and Rust proposed that the intraclass correlation for stratum h, ρ_{hy}, be estimated by

$$\hat{\rho}_{hy} = \frac{deff_h - \left\{1 + cv_h^2(w)\right\}}{\left\{1 + cv_h^2(w)\right\} \times (b_h^* - 1)}, \tag{5}$$

where $deff_h$ is the estimated design effect of the estimator in stratum h. [Equation (5) corrects a typo in (4) in Chen and Rust]. The averages of the estimates of ρ_{hy} reported below for several outcome variables were obtained using this method.

Table 2 presents estimates of the averages of the within-domain intraclass correlations for a selection of variables collected in PHIA, particularly variables used in producing HIV-related estimates. For ease of presentation, and because of the general similarities of the $\hat{\rho}_{hy}$, the values of $\hat{\rho}_y$ in the table are the averages of the within-domain $\hat{\rho}_{hy}$ values, averaged across both domains and countries.

The first four estimates in table 2 are key estimates for the PHIA surveys. The average $\hat{\rho}_y$ values for these variables are low; moreover, the $\hat{\rho}_y$ values for the ART-related estimates are based on subclasses, an issue taken up in the next section. The estimates of the intraclass correlations for some questionnaire items, such as high school attendance and receipt of economic support, are higher, as might be expected for some socio-economic characteristics.

4.3 Values of the Relative Precision Measure $\widehat{R}$

As can be seen from (4), the values of $\hat{R}$ depend on $cv^2(K)$, the desired subsample size within sampled PSUs b^*, and the average domain $\hat{\rho}_y$ s for key survey variables. The latter two quantities are affected by the choice of sample design. The sample designer determines the subsample size taking into account the survey's fieldwork plan, and different surveys may well choose different subsample sizes. The within-stratum intraclass correlations are lower than the overall intraclass correlations when the strata comprise PSUs that have similar mean values for key survey variables. The correlations ρ_y depend on both the survey variables and the stratification used in the design.

For purposes of illustration, we assume that $b^* = 50$ for estimates based on the full sample. This value is larger than the within PSU sample size used in all the analyzed PHIA surveys

except for country A. For cross-class estimates (subclasses that are fairly evenly distributed across the PSUs), the b^* s are based on the subclass sample sizes. For example, for estimates for men and women separately, $b^* < 25$; for persons who are HIV positive, b^* varies between five and eight across countries; and for those on ART, b^* is around five or less. Table 3 displays values of the relative magnitude of the sample sizes for the FSS design to that for the SW design to attain the same level of precision for the survey estimates–that is, measured by $\widehat{R}$ in (4)–for two values of $cv^2(K)$ and for various values of b^* and $\widehat{\rho}_y$.

The findings in table 3 are as expected. The overall design effect for the FSS design includes the factor $deff_w = \left\{1 + cv^2(K)\right\}$ that is absent in the overall design effect for the SW design. Thus, the larger $cv^2(K)$, the greater is the value of $\widehat{R}$, reflecting the relative lower precision of the FSS design. The clustering component of the overall design effect is smaller when b^* is small, as for small cross-classes, and for smaller values of $\widehat{\rho}_y$. Hence the value of $\widehat{R}$ is larger in such cases. In the PHIA surveys, ρ_y values are low for the key survey variables, and cross-class estimates are of great importance. In this situation, the value of $\widehat{R}$ approaches $1 + cv^2(w) = 1 + cv^2(K)$ for some estimates.

5. GENERAL CONSIDERATIONS

The FSS design has benefits of more straightforward fieldwork operations as compared with the EP and SW designs. The FSS design is widely used in the Demographic and Health Surveys (ICF International 2012) and in many of UNICEF's Multiple Indicator Cluster Surveys (see UNICEF 2020). The sampling manual produced by the United Nations Department of Economic and Social Affairs (2008, p. 72) recommends the FSS design over the EP design. However, as shown in section 2, the EP and SW designs are more statistically efficient than the FSS design; that is, the FSS design requires a larger sample size than the EP and SW designs to produce estimates with the same levels of precision. For this reason, the SW design has been used in most countries where the PHIA surveys have been conducted.

With the FSS design, the sample designer first needs to decide on the subsample size in each sampled PSU (e.g., twenty-five households) and then determine how many PSUs to select. This latter task is not as simple as it might appear: to produce survey estimates of prescribed levels of precision, the weighting design effect $\left\{1 + cv^2(w)\right\}$ should be factored into the calculation of the variances and that quantity is difficult to determine prior to listing.

To achieve an approximation to the specified overall sample size with the widely used EP design, the sample designer needs to obtain an external estimates of the population sizes N for any domains for which specified levels of precision have been set, and estimate the noncoverage rate. In contrast, the two-stage two-phase SW design used in the PHIA surveys produces an SW sample of the desired size without the need to estimate these quantities. Both the EP and SW designs result in variable subsample sizes in the sampled PSUs. This variation in subsample sizes presents some operational fieldwork challenges, but with well-managed field organization, these challenges have been successfully overcome in the PHIA surveys. A concern that unequal workloads could affect response rates did not

materialize. Some excessively large subsample sizes occurred in a few of the PHIA surveys, and they were capped, with weighting adjustments made in compensation. These weighting adjustments give rise to the weighting design effect $deff_w$ in formula (3); however, with capping of the subsample sizes rarely implemented, and with care taken to ensure that the weight adjustments from capping are small, the weighting design effect was only about 1.02 in those PHIA surveys where capping was employed.

The FS, EP, and SW sample designs must make allowance for the effect of nonresponse on achieved sample sizes, which can be handled in each case by increasing the expected number of dwellings to sample in each PSU. With all three designs, lists of dwelling units are compiled in each sampled PSU. With the FSS and EP designs, the sample dwellings could be selected on a flow basis, from one PSU to the next, whereas with the SW design, the selections can be made only after the listings have been completed for a domain or the full country. The EP and SW designs can employ a liberal definition of listing units to be sampled because sampled units found not to contain households are simply dropped from the sample as ineligible at the time of data collection. For instance, dwellings that are unoccupied at the time of listing (including dwellings under construction) may be given a chance of selection because they may have become occupied by the time of data collection. The FSS design uses a more conservative definition to achieve its specified household sample size in each PSU.

The FSS and the EP designs could, in theory, achieve cost savings by combining the listing and data collection operations. Thus, the survey team would visit the PSU for listing the dwelling units, draw a sample of them, and continue to data collection. In practice, that procedure is generally not favored because of concerns about its effect on the quality of the listing and sampling operations. Often these operations are similar for the three designs with no appreciable differences in costs. Dealing with the unequal subsample sizes by PSU with the EP or SW designs is more complex, but it is not clear that it has a significant impact on fieldwork costs.

The main benefit of the EP and SW designs over the FSS design is the smaller sample size needed to achieve a given level of precision for the survey estimates. This benefit is of particular importance for the PHIA surveys because of their expensive data collection operations that include blood draws, shipping of blood samples for processing, and processing costs. For example, with a $cv^2(K)$ of 0.2, the FSS design requires a sample size that is almost 20 percent larger than that with the SW design. Since the subsample size b has been chosen to satisfy logistical and efficiency specifications, the natural way to achieve the 20 percent increase in effective sample size is to increase the number of sampled PSUs by 20 percent. If instead the subsample sizes were increased within selected PSUs, the resulting increase in the clustering design effect, $deff_c$, would lead to the need for an increase of more than 20 percent in the sample size to attain the 20 percent increase in effective sample size.

The best solution to the problems created by inaccuracies in the measures of size is to develop better size measures for use in the PPES selection. In this regard, it is worth noting that subjectively modifying some PSU measures of size does not harm the integrity of the sample, but it can sometimes improve the sample efficiency. For example, the census-based

measures of size could be doubled, say, in PSUs on the outskirts of large towns where major growth is likely to have occurred. This kind of modification can be useful if it produces measures of size that are better aligned with the current PSU sizes.

Acknowledgments

The authors thank the editors, the reviewers, and Westat colleagues Mike Brick, Adam Chu, Jean Opsomer, and Keith Rust for their helpful comments.

The Population-based HIV Impact Assessment (PHIA) project is supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (CDC). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the funding agencies.

REFERENCES

Chen S, and Rust K (2017), "An Extension of Kish's Formula for Design Effects to Two- and Three-Stage Designs with Stratification," Journal of Survey Statistics and Methodology, 5, 111–130. [PubMed: 37583392]

Fuller WA (2009), Sampling Statistics, Hoboken, NJ: Wiley.

Gabler S, Haeder S, and Lahiri P (1999), "A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering," Survey Methodology, 25, 105–106.

Holt D (1980), "Discussion of 'Sample Designs and Sampling Errors for the World Fertility Survey' by Verma, V., Scott, C. and O'Muircheartaigh, C," Journal of the Royal Statistical Society, A, 143, 468–469.

International ICF (2012), Demographic and Health Survey Sampling and Household Listing Manual. Measure DHS. Calverton, MD: ICF International. Available at https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf. Accessed July 6, 2020.

Kish L (1965), Survey Sampling, New York: Wiley.

————. (1987), "Weighting in Deft^2," Survey Statistician, June. International Association of Survey Statisticians, pp. 26–30.

————. (1992), "Weighting for Unequal P_i.," Journal of Official Statistics, 8, 183–200.

————. (1995), "Methods for Design Effects," Journal of Official Statistics, 11, 55–77.

Särndal C-E, Swensson B, and Wretman J (1992), Model Assisted Survey Sampling, New York: Springer-Verlag.

Skinner C (1986), "Design Effects of Two-Stage Sampling," Journal of the Royal Statistical Society, B, 48, 89–99.

Spencer BD (2000). "An Approximate Design Effect for Unequal Weighting When Measurements May Correlate with Selection Probabilities," Survey Methodology, 26, 137–138.

UNICEF (2020), MICS Surveys. Available at https://mics.unicef.org/. Accessed July 6, 2020.

United Nations Department of Economic and Social Affairs (2008), Designing Household Survey Samples: Practical Guidelines. Studies in Methods, Series F, No. 98. New York: United Nations. Available at https://unstats.un.org/unsd/demographic/sources/surveys/Handbook23June05.pdf. Accessed July 6, 2020.

Valliant R, Dever JA, and Kreuter F (2013), Practical Tools for Designing and Weighting Survey Samples, New York: Springer.

Table 1.

Values of National Average Domain Values of $\overline{K}_c$ and of $cv^2(K)$ for Thirteen PHIA Surveys

Country	Years since the last census	Ave. $\overline{K}_c$	Ave. $cv^2(K)$
A	5	1.34	0.46
B	10	1.67	0.36
C	6	1.41	0.35
D	3	1.28	0.32
E	4	1.51	0.27
F	7	1.42	0.21
G	9	1.22	0.20
H	5	1.24	0.20
I	3	1.20	0.12
J	2	1.40	0.10
K	1	1.08	0.08
L	4	1.08	0.07
M	6	1.26	0.06

Table 2.

Average Domain Intraclass Correlation Estimates ($\hat{\rho}_y$) Averaged across Countries for Selected PHIA Variables, Persons 15–49 Years Old

Estimate (% of persons with the characteristic)	Average $\hat{\rho}_y$
HIV positive	0.02
Ever tested for HIV	0.03
On ART among those who tested positive for HIV	0.04[a]
Viral load suppression among those on ART	0.02[a]
Paid work in the past twelve months	0.03
Ever attended school	0.03
Has attended high school (eighteen years of age and older)	0.10
Lives in a household that has received economic support in the past year	0.10

[a]These estimates are based on small sample sizes and are less reliable.

ART, antiretroviral therapy.

Table 3.

Values of the Ratio $\widehat{R}$ of $deff$ with the FSS Design to $deff$ with the SW Design for Various Values of b^* and $\widehat{\rho}_y$ and for Two Values of $cv^2(K)$

(a) $cv^2(K) = 0.25$

b^*	$\widehat{\rho}_y$			
	0.01	0.03	0.05	0.10
50	1.17	1.10	1.07	1.04
30	1.19	1.13	1.10	1.06
20	1.21	1.15	1.12	1.08
10	1.23	1.19	1.16	1.12
5	1.24	1.22	1.20	1.16

(b) $cv^2(K) = 0.10$

b^*	$\widehat{\rho}_y$			
	0.01	0.03	0.05	0.10
50	1.07	1.04	1.03	1.02
30	1.08	1.05	1.04	1.02
20	1.08	1.06	1.05	1.03
10	1.09	1.08	1.07	1.05
5	1.10	1.09	1.08	1.06