Dealing with Inaccurate Measures of Size in Two-Stage Probability Proportional to Size Sample Designs: Applications in African Household Surveys

Graham Kalton[1], Ismael Flores Cervantes[1], Carlos Arieira[1],
Mike Kwanisai[1], Elizabeth Radin[2,3], Suzue Saito[2,3], Anindya De[4†],
Stephen McCracken[4†], and Paul Stupp[5]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850
[2]ICAP at Columbia University, New York, NY 10032
[3]Department of Epidemiology, Mailman School of Public Health of Columbia University, New York NY 10032
[4]U.S. Centers for Disease Control and Prevention, Atlanta, GA 30333
[5]U.S. Centers for Disease Control and Prevention, retiree

Abstract
Two-stage sample designs are used for household surveys in many countries. At the first stage, primary sampling units (PSUs) are sampled with probabilities proportional to their estimated sizes (PPES). A list of households is compiled in the selected PSUs, and households are selected with equal probability from each PSU. With this design, an overall equal probability sample design would yield a constant number of households from each sampled PSU if the measure of size used in the PPES selection were directly proportional to the number of households listed. However, there are often sizable differences between the measures of size used in the PPES selection and the listed sizes. Two common methods for dealing with these differences are: (1) to retain the equal probability sample design, allowing the sample size to vary across the sampled PSUs; and (2) to retain the fixed sample size in each PSU and to compensate for the unequal selection probabilities by weighting. This paper discusses the theoretical and practical advantages and disadvantages of these two methods. The discussion is illustrated with data from the Population-based HIV Impact Assessment (PHIA) surveys that have been conducted in several African countries. In all of these countries the PSUs were the enumeration areas (EAs) used in the most recent population census, and they were sampled with probabilities proportional to the EAs' population sizes at the time of the census.

Key Words: design effect, clustering effect, weighting effect, equal probability sample, equal subsample size

1. Introduction

Stratified two-stage sample designs are used for household surveys in many countries, where the main strata are often geographic subdivisions such as provinces or administrative regions. Within each stratum, the primary sampling units (PSUs) are sampled with

---

[†]Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention/the Agency for Toxic Substance and Disease Registry.

probabilities proportional to size (PPS) or, more realistically, with probabilities proportional to estimated size (PPES), where the measures of size are estimates of such quantities as the numbers of households or persons in the PSU. In many African countries, the PSUs are census enumeration areas, and the measures of size are based on data collected in the last population census. In preparation for the second stage of sampling, lists of households are compiled for the selected PSUs. When a sampled PSU is very large so that listing the whole PSU would be problematic, the PSU may be divided into segments, with one segment being selected by PPES for listing (see Section 3 for further details). Samples of households are then selected from the lists, generally by systematic sampling. Either all or a sample of the households' residents are then included in the survey. Since the last census has often been carried out several years earlier, the measures of size are sometimes seriously inaccurate measures of the PSUs' current sizes. This paper compares two common methods for dealing with this inaccuracy.

If an exact PPS sample were selected within a given first-stage sampling stratum, the probability of household β in PSU α being included in the sample would be

$$P(\alpha\beta) = \left(\frac{aN_\alpha}{\Sigma N_\alpha}\right)\left(\frac{b_\alpha}{N_\alpha}\right) = \left(\frac{ab_\alpha}{\Sigma N_\alpha}\right) \tag{1}$$

where a PSUs are selected 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 ideal sample design is statistically efficient. It is also operationally attractive because it produces the same interviewing workload in each sampled PSU and the total sample size is predetermined. In practice, however, this ideal 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

$$P(\alpha\beta) = \left(\frac{aM_\alpha}{\Sigma M_\alpha}\right)\left(\frac{b_\alpha}{N_\alpha}\right) = \left(\frac{a}{\Sigma M_\alpha}\right)\left(\frac{b_\alpha M_\alpha}{N_\alpha}\right) \tag{2}$$

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., 25 households) from each selected PSU in each stratum (the fixed take may vary between strata). With this FSS design, households are sampled with unequal probabilities that require weighting compensation in the analyses.
- The second approach, termed the equal probability (EP) design, allows the sample sizes to vary across the sampled PSUs in order to obtain an overall equal probability (*epsem*) design within each stratum.

The FSS design, which is widely used in the Demographic and Health Surveys (ICF International, 2012) and UNICEF's Multiple Indicator Surveys, has benefits for fieldwork operations. The sampling manual produced by the United Nation 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 design is more statistically efficient than the FSS design. As a result, the FSS design requires a larger sample size than the EP design to produce estimates with the same levels of precision. For this reason, the EP design has been used (with minor modifications) in the Population-based HIV Impact Assessment (PHIA) surveys in most of the countries in which the surveys have been conducted.

The paper is organized as follows. Section 2 presents the theory for the comparison of the statistical efficiencies of the two 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 and, to a lesser extent, on the homogeneity of the variable under study for a particular analysis within the PSUs, measured by the intraclass correlation. Section 3 describes the sample designs used in the PHIA surveys for which the relative efficiencies of the FSS and EP have been computed. Section 4 reports results for the components affecting the relative precision of the EP and FSS 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

As background, let the probability of selecting household β in PSU α for the sample in a given stratum be denoted by

$$P(\alpha\beta) = \left(\frac{aM_\alpha}{\Sigma M_\alpha}\right)\left(\frac{b_\alpha}{N_\alpha}\right) = \frac{ab_\alpha}{MK_\alpha}, \tag{3}$$

where a PSUs are selected at the first stage with probabilities proportional to estimated sizes M_α, $M = \Sigma M_\alpha$, b_α households are selected with equal probability at the second stage from the N_α households listed for the survey, and $K_\alpha = N_\alpha / M_\alpha$. If $K_\alpha = K$ and $b_\alpha = b,$ then the overall selection probability $P(\alpha\beta) = f$, a constant. In this case, the overall sample size is $n = Kab$ and, if $K = 1$ the overall sample size is $n = ab.$

In general, K_α varies across the sampled PSUs. With the FSS design for selecting households within sampled PSUs, and with epsem sampling of households within PSUs, a household's overall selection probability is given by

$$P(\alpha\beta) = \frac{ab}{MK_\alpha}. \tag{4}$$

With this unequal probability sample design for sampling households, weights proportional to K_α are needed in the analysis. These weights decrease the precision of the survey estimates.

The EP sample design avoids the need for weights in the analyses. However, that benefit comes at the price of variability in the sample sizes in the sampled PSUs, with the sample size in sampled PSU α being proportional to K_α. This variability in b_α across the sampled PSUs presents fieldwork challenges and sometimes difficulties fixing the overall sample size.

We assess the relative precision of survey estimates obtained under these two designs by comparing the design effects of sample means for the two designs. Two factors contribute to these design effects: clustering and weighting. With equal-sized PSUs (clusters) and an overall epsem design, the clustering design effect is $Deff_c = 1 + (b-1)\rho,$ where b is the sample size in each PSU and ρ is the intraclass correlation for a given variable within the PSUs. With an overall epsem design but variable $b_\alpha,$ b is often replaced by $\bar{b}$ (see, for example, Kish, 1965), but Holt (1980) and Skinner (1986) show that it is better replaced by $\Sigma b_\alpha^2 / \Sigma b_\alpha = \bar{b}[1 + cv^2(b)],$ where $cv(b)$ denotes the coefficient of variation of the $b_\alpha.$ (Note that throughout this paper, $cv^2(x) = \text{var}(x) / \bar{x}^2$ with $\text{var}(x) = \Sigma(x_i - \bar{x})^2 / n$ 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 weight (Kish, 1992). This assumption does not always hold, but it is generally a reasonable approximation in the current context. Kish (1987) proposed estimating the overall design effect for a sample mean with an unstratified two-stage design and variable weights as the product of $Deff_c$ and $Deff_w,$ i.e., by

$$Deff = \{1 + cv^2(w)\} \times \{1 + (\bar{b} - 1)\rho\}, \tag{5}$$

In a model-based justification of Kish's formula, Gabler, Haeder, and Lahiri (1999) developed a modification of $\bar{b},$ replacing it by

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

where $w_{\alpha\beta}$ is the weight for household β in PSU α. With an equal probability sample of households within each sampled PSU, as assumed throughout this paper, $w_{\alpha\beta} = w_\alpha$ $w_\alpha = K_\alpha \Sigma M_\alpha / ab_\alpha,$

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

$$1 + cv^2(w) = 1 + \frac{\text{var}(w)}{\bar{w}^2} = \frac{n \Sigma b_\alpha w_\alpha^2}{(\Sigma b_\alpha w_\alpha)^2},$$

where $n = \Sigma b_\alpha,$ $\bar{w} = \Sigma b_\alpha w_\alpha / n,$ and $\text{var}(w) = \Sigma b_\alpha (w_\alpha - \bar{w})^2 / n.$

Consider any sample allocation of the b_α at the second stage subject to fixed $n = \Sigma b_\alpha$. Then $w_\alpha \propto K_\alpha / b_\alpha$. Hence

$$1 + cv^2(w) == n(\Sigma K_\alpha^2 / b_\alpha) / (\Sigma K_\alpha)^2$$
$$\{1 + (n*-1)\rho\} = \{1 + [\Sigma K_\alpha^2 / \Sigma(K_\alpha^2 / b_\alpha) - 1]\rho\}$$

and the overall design effect is

$$Deff = \frac{n(1-\rho)\Sigma(K_\alpha^2 / b_\alpha)}{(\Sigma K_\alpha)^2} + n\rho \frac{\Sigma K_\alpha^2}{(\Sigma K_\alpha)^2}$$

$Deff$ is minimized when $\Sigma K_\alpha^2 / b_\alpha$ is minimized subject to a fixed overall sample size $n = \Sigma b_\alpha$. The Cauchy inequality states that $(\Sigma a_h^2)(\Sigma b_h^2) \geq (\Sigma a_h b_h)^2$, with the minimum occurring when $a_h / b_h = C$, a constant. Applying this inequality to $(\Sigma K_\alpha^2 / b_\alpha)(\Sigma b_\alpha)$ gives a minimum value when $(K_\alpha b_\alpha^{-1/2}) / b_\alpha^{1/2} = C$, i.e., $b_\alpha \propto K_\alpha$. Thus, the EP design with $b_\alpha \propto K_\alpha$ minimizes $Deff$ and hence is the most efficient design.

We turn now to the relative efficiency of the FSS and EP designs. With the FSS design, $b^* = b$ and $1 + cv^2(w) = \Sigma K_\alpha^2 / (\Sigma K_\alpha)^2$. With the EP design, $w_\alpha = w$, $\{1 + cv^2(w)\} = 1$, and b^* reduces to $\Sigma b_\alpha^2 / \Sigma b_\alpha$, as derived by Holt (1980) and Skinner (1986). With the EP design, $b_\alpha = nK_\alpha / \Sigma K_\alpha$, and hence

$$\begin{aligned} b^* &= [n^2 \Sigma K_\alpha^2 / (\Sigma K_\alpha)^2] / [n\Sigma K_\alpha / \Sigma K_\alpha] = n\Sigma K_\alpha^2 / \Sigma K_\alpha \\ &= n\Sigma K_\alpha [1 + cv^2(K)] \\ &= \bar{b}\bar{K}[1 + cv^2(K)] \end{aligned}$$

where $\bar{b} = n / a$ and $\bar{K} = \Sigma K_\alpha / a$. Note that the term $\{1 + cv^2(K)\}$ in the EP design is equal in magnitude to $F = \{1 + cv^2(w)\}$ that applies with the FSS design. The overall sample size with the EP design is equal to the planned sample size only if $\bar{K} = 1$ for the selected sample of PSUs. When this is not the case, the planned sample size can be achieved by setting the value of $\bar{b}$ to b once the PSUs have been selected and the counts N_α have been determined for the selected PSUs.

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

$$\begin{aligned} Deff_{fss} &= F\{1 + (b-1)\rho\} \\ Deff_{ep} &= \{1 + (Fb-1)\rho\} \end{aligned}$$

assuming that $\bar{b} = b$ with the EP design. Hence

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

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

$$R = \frac{Deff_{fss}}{Deff_{ep}} = 1 + \frac{(F-1)(1-\rho)}{1+(Fb-1)\rho} = 1 + \frac{cv^2(K)(1-\rho)}{1+(b^*-1)\rho} \tag{6}$$

To attain the same precision, the sample size for the FSS design needs to be larger than that for the EP design by this ratio of design effects, R. In particular, this ratio will be relatively large when ρ is small (say, 0.05 or even 0.10), b is small, and the variability in the K_α is large.

Chen and Rust (2017) have extended the general design effect model given by equation (5) to two- and three-stage designs with stratification. They propose that the intraclass correlation for stratum *h*, ρ_h, be estimated based on equation (5) as

$$\hat{\rho}_h = \frac{deff_h - \{1 + cv_h^2(w)\}}{\{1 + cv_h^2(w)\} \times (b_h^* - 1)} \tag{7}$$

where $deff_h$ is the estimated design effect of the estimator in stratum *h*. (Equation 7 corrects a typo in equation (4) in Chen and Rust.)

3. The Sample Designs for the PHIA Surveys

The sample designs of each of the thirteen PHIA surveys covered in this paper were 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. To satisfy the second objective, the overall sample was stratified by relevant subdivisons of the country (such as region or province), and the total sample size was allocated to the strata 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 stratum the stratum. 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 of a specified age range in the selected households were included in the study sample. Children were subsampled, as were adults for some special studies. This paper deals only with the adult population of primary interest, that is, persons between the ages of 15 and 49 years old.

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, a third stage of sampling was introduced: the large sampled PSUs were divided into segments, one segment was selected in each PSU, and the listing operation was applied only in the sampled segment. Let the estimated number of households in segment β in PSU α at the time of the census be $\hat{M}_{\alpha\beta} = p_{\alpha\beta} M_\alpha$ where $p_{\alpha\beta}$ is estimated from the listing operation (with $\Sigma_\beta p_{\alpha\beta} = 1$ and hence $\Sigma_\beta \hat{M}_{\alpha\beta} = M_\alpha$). Then the selection of segment β within PSU α with a probability of $\hat{M}_{\alpha\beta} / M_\alpha$ gives an overall selection probability for that segment of $a\hat{M}_{\alpha\beta} / M$. Hence, with a single segment being sampled from each segmented PSU, the design can still be treated as a two-stage sample with the segments as PSUs that are sampled with probabilities proportional to the estimated census counts $\hat{M}_{\alpha\beta}$. In some cases, the PSU was divided into d_α segments of roughly equal size, and one was sampled with probability $1/d_\alpha$.

4. Relative Precision *R* Based on Several PHIA Surveys

It can be seen from Equation (6) that the magnitude of the ratio of the design effects R depends on the values of the two parameters $cv^2(K)$ and ρ, and on the average sample size per PSU chosen for the sample design. The next two subsections report estimates of these two parameters for a number of PHIA surveys, and the last subsection then examines the values of R for a variety of survey estimates of population percentages in the PHIA surveys.

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

Estimates of $cv^2(K)$ and $\bar{K}$ are useful for planning a sample design. The value of $cv^2(K) = cv^2(w)$ is of major importance for determining the overall sample size for the FSS design so the survey estimates will meet specified precision levels but it has only a minor effect on the sample size for the EP design (through b^*). The value of $\bar{K}$ affects the achieved sample size with the EP design (unless the overall sampling fraction is based on the completed listings in the sampled PSUs) but it has no impact on sample size with the FSS design.

The values of the national averages of $cv^2(K)$ were calculated separately for regional (or provincial) domains of each of the 13 countries and then averaged across the domains. The averaging was done in a way that retained the domain stratification but removed the PHIA disproportionate allocation across domains; thus, the estimates of the $cv^2(K)$ are applicable for a design with a proportionate allocation with the same strata. The results are displayed in the fourth column of Table 1, with the countries listed 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 might be expected, the $cv^2(K)$ are in the lower part of the range for I, J, and K, the three countries that had censuses within three years of the survey. It might be hypothesized that the $cv^2(K)$'s would be larger in countries where more segmentation was applied, both because of the error introduced by the need to estimate $\hat{M}_{\alpha\beta}$ and by the fact that abnormal growth was one reason for the need to employ segmentation. However, Table 1 presents no evidence to support this hypothesis.

The final column of Table 1 presents the national averages of the within domain K_α (with the effect the disproportionate stratification removed). These averages reflect both the growth in the number of households since the last census and also the fact that the numerator includes listed units that were unoccupied at the time of survey data collection. Another difference between the two counts is that some of the listed units may contain more than one household.

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 $cv^2(K_h)$

and $\bar{K}_h$ are therefore important. Both these quantities vary markedly across regions. For example, the $cv^2(K_h)$'s for country B range from 0.07 to 0.48; for country F they range from 0.13 to 0.33; and for country L they range from 0.03 to 0.15. The ranges of the regional $\bar{K}_h$'s for these countries were as follows: 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.

Table 1: Values of national average values of $cv^2(K)$ and of $\bar{K}$ calculated from 13 PHIA surveys

Country	*Years since last census*	*Segmented PSUs (%)*	*Ave.* $cv^2(K)$	*Ave.* $\bar{K}$
A	5	1	0.60	1.35
B	10	20	0.36	1.67
C	6	5	0.35	1.41
D	3	14	0.32	1.28
E	4	28	0.27	1.51
F	7	55	0.21	1.42
G	9	27	0.20	1.22
H	5	28	0.20	1.24
I	3	13	0.12	1.20
J	2	0	0.10	1.40
K	1	13	0.08	1.08
L	4	0	0.07	1.08
M	6	2	0.06	1.26

4.2 Values of the Intraclass Correlations ρ

As can be seen in equation (6), the value of R depends on the intraclass correlation coefficient ρ for the particular variable under study. To examine the magnitude of the intraclass correlations, regional level estimates of $\hat{\rho}_h$ were computed using equation (7) for a number 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 $\hat{\rho}$ values for a given variable across countries, the $\hat{\rho}$ have been averaged across countries. These averages are presented in Table 2.

Table 2: Averages of estimates of within-strata intraclass correlations $\hat{\rho}$ across countries for selected PHIA variables, persons 15-49 years old.

Estimate (% of persons with the characteristic)	*Average* $\hat{\rho}$
HIV positive	0.02
Ever tested for HIV	0.03
On antiretroviral therapy (ART) among those who tested positive for HIV	0.04†
Viral load suppression among those on ART	0.02†
Paid work in the past 12 months	0.03
Ever attended school	0.03
Has attended high school (18 years of age and older)	0.10
Lives in a household that has received economic support in the past year	0.10

†These estimates are based on small sample sizes and are less reliable.

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

4.3 Values of the Relative Precision Measure *R*

As can be seen from equation (6), values of R depend on $cv^2(K)$, and also on the average ρ_h and the desired subsample size within sampled PSUs. The latter two quantities are affected by the choice of sample design. The within-stratum intraclass correlation ρ_h depends on the stratification used in the design, and the sample designer chooses the desired subsample size taking account of the fieldwork plan. For purposes of illustration, we assume that $b^* = 50$ for full sample estimates. This value is larger than that used in all the PHIA surveys analyzed except for country A. For cross-class estimates (subclasses that are fairly evenly distributed across the PSUs), the b^*'s are 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 for R computed using equation (6) for two values of $cv^2(K)$ and for various values of b^* and ρ.

Table 3: Values of R for various values of b^* and ρ for two values of $cv^2(K)$.

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

b^*	ρ			
	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^*	ρ			
	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

The findings in Table 3 are as expected. The overall design effect for the FSS design includes a factor for $Deff_w = 1 + cv^2(K)$ that is absent in the overall design effect for the EP design. Thus, the larger $cv^2(K)$, the greater is the value of R, reflecting the relative lower precision of the FSS design. The value of R is greater when the value of b^* is small, as for small cross-classes. It is also greater for smaller values of ρ. These two factors determine $Deff_c = 1 + (b^* - 1)\rho$. In the PHIA surveys, ρ values are low for the key variables and cross-class estimates are of great importance. In this situation, the value of R approaches $1 + cv^2(K)$ for some estimates.

5. Concluding Remarks

The FSS design has the attraction of operational simplicity. The sample designer needs only to determine the subsample size in a PSU (e.g., 25 households) and then determine how many PSUs to select. This latter task is, however, not as simple as it might appear: in order to produce estimates of prescribed precision, the weighting design effect $\{1 + cv^2(w)\}$ should be factored into the calculation, and that term is difficult to determine prior to listing. The main attraction of the FSS design relates to the fieldwork, where equal interviewer loads across all PSU makes fieldwork organization very straightforward.

In contrast, the EP design presents two challenges. To achieve a specified sample size, the sample designer needs to know $\bar{K}$ for the full sample and also for any domains for which specified levels of precision have been set. In the PHIA surveys, this issue has been resolved by carrying out the listings for all sampled PSUs prior to selecting any households. Thus, $\bar{K}$ is known and can be used in determining the required sampling fraction to yield the desired sample size. The EP sample design produces unequal subsample sizes in the sampled PSUs. This variation in subsample sizes presents some operational challenges but, with well-managed field organization, this challenge has been successfully overcome in the PHIA surveys. A concern that unequal workloads could affect response rates has not materialized. In a few cases, excessively large subsample sizes have been capped, with weighting adjustments made in compensation.

In choosing between the EP and FSS designs, the smaller sample size needed to achieve a given level of precision for survey estimates with the former design has to be balanced against the operational simplicity of the latter design. For the PHIA surveys with their key HIV-related estimates, it was decided that the complexity of the EP design was preferred over a substantial increase in sample size. For example, with a $cv^2(K)$ of 0.2, the sample size with the FSS design would need to be almost 20 percent larger, i.e., a 20 percent increase in the number of PSUs sampled, to give the same precision as the EP design.

The best solution to inaccuracies in the measures of size would be to develop better size measures to be used in the PSU selection. In this regard, it is worth noting that modifying some measures of size subjectively will not harm the integrity of the sample. For example, the census-based measures of size could be doubled, say, in PSUs on the outskirts of large towns where major growth is expected to have occurred. The only issue is whether the new measures are more closely aligned with the current sizes.

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.

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.

ICF International (2012). *Demographic and Health Survey Sampling and Household Listing Manual*. MEASURE DHS. Calverton, MD: ICF International. https://dhsprogram.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf

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

Kish, L. (1987). Weighting in Deft^2. *Survey Statistician*, June, pp. 26-30. International Association of Survey Statisticians.

Kish, L. (1992). Weighting for unequal P_i. *Journa of Official Statistics*, **8**, 183-200.

Skinner, C. (1986). Design effects of two-stage sampling. *Journal of the Royal Statistical Society, B*, **48**, 89-99.

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. https://unstats.un.org/unsd/demographic/sources/surveys/Handbook23June05.pdf