Review of Income and Wealth
Series 44, Number 1, March 1998

DOES SAMPLE DESIGN MATTER FOR POVERTY RATE COMPARISONS?

BY STEPHEN HOWES AND JEAN OLSON LANJOUW

World Bank and Yale University

Poverty comparisons—an increasingly important starting-point for welfare policy analysis—are almost always based on household surveys. Therefore they require that one be able to distinguish underlying differences in the populations being compared from sampling variation: standard errors must be calculated. This has typically been done assuming that the household surveys are simple random samples. However, household surveys are more complex than this. We show that taking into account sampling design has a major effect on estimated standard errors for well-known poverty measures. In our samples they increase by around one-half. We also show that making only a partial correction for sample design (taking into account clustering, but not stratification, whether explicit or implicit) can be as misleading as not taking any account of sampling design at all.

1. INTRODUCTION

Has poverty increased or fallen? Is urban or rural poverty higher? Will some proposed policy reduce or increase poverty? These are typical of the questions asked in poverty analyses. To provide answers, recourse to household surveys is required. However, surveys are not censuses. They are samples, with a size typically numbering in the thousands of households from which conclusions concerning populations typically numbering in the millions must be drawn. Thus any comparative analysis must distinguish population differences from sampling variation. A series of recent papers have stressed the importance of this and have provided the tools by which standard errors can be calculated (Howes, 1993; Kakwani, 1993; Pudney and Sutherland, 1994; Ravallion, 1994). The problem with these papers is that, in presenting statistical methods and results for use in poverty comparisons, they make the assumption that the household surveys being analyzed are simple random samples of the populations from which they are drawn. In fact, however, they are not. Household surveys are typically far more complex in their design, incorporating stratification and clustering.

What are the implications for poverty analysis of complex survey design? Two main biases are introduced by ignoring sample design in calculating standard errors. On the one hand, ignoring stratification will lead to an *over*estimate of standard errors. This is because, with stratification, one ensures that no part of the sampling frame goes unrepresented. On the other hand, ignoring clustering will lead to an *under*estimate of standard errors if, as is typically the case, there is homogeneity within the cluster with respect to the variable of interest. The

Note: We would like to thank, for their provision of data, information and/or comments: Benu Bidani, Gaurav Datt, Mark Foley, Margaret Grosh, Dean Joliffe, Peter Lanjouw, Martin Ravallion, Christopher Scott, Kinnon Scott, Salman Zaidi, and Qinghua Zhao. Mr Ranzam of Pakistan's Federal Bureau of Statistics also provided us with useful information.

intuition is the same as that for stratification: clustering leads to a less uniform coverage of the population, and so reduces precision. Since these biases run in opposite directions, we explore this issue using actual household surveys. In the two examples we provide, using formulae which assume simple random sampling leads to estimates of standard errors for poverty measures which are only two-thirds the size they should be. That is, ignoring sample design can make us think that our estimates are substantially more precise than they actually are.

The next section describes the types of poverty measures we will be discussing. Section 3 sets out assumptions which characterize the design of many household surveys and then provides formulae for calculating standard errors which are consistent with these assumptions. Section 4 gives the two examples, and Section 5 concludes.

2. Poverty Measures

In this section, we consider the class of additive poverty measures. Although this class excludes some poverty measures, such as the Sen index, it includes many well-known measures, such as the head-count and the poverty gap. In addition to being widely used, comparisons of these two measures, over varying poverty lines, form the basis of, respectively, first and second-order stochastic dominance analysis (see Atkinson, 1987). Let $\pi_j = \pi(k_j)$ be a measure of household poverty for the jth household, where $k_j = f(y_j, X_j)$. y_j is total household income/consumption and X_j is a vector of variables such as prices and household size which allow one to normalize y_j such that households with equal k's are equally well off. In the simplest case, k_j will be per capita nominal income, y_j/h_j, where h_j is household size.

A functional form for $\pi(\cdot)$ is based on the indicator function, $I(k_j \leq k)$, where k is the poverty line. If the expression inside the brackets is false then $I(\cdot)$ is zero. If it is true then $I(\cdot)$ equals one if one is interested in poverty among households; it equals h_j if one interested in the number of poor individuals.

We can now define specific poverty measures. For the head-count ratio, $\pi(k_j) = I(k_j \leq k)$. For the poverty gap, $\pi(k_j) = I(k_j \leq k) * (1 - k_j/k)$. More generally, the well-known and much-used FGT family is given by $\pi_\alpha(k_j) = I(k_j \leq k) * (1 - k_j/k)^\alpha$ for $\alpha \geq 0$ (see Ravallion, 1994, for further details).

The next step is to aggregate the π_j into a summary poverty index. Define π as an estimator of this index. It is equal to an estimator, t, of "total poverty" (e.g. in the case of the head-count ratio, the total number of poor people or households), normalized by an estimator, p, of population size. As π is a ratio of two random variables, it is biased but consistent.

The complexity of survey design is very easily incorporated into the calculation of unbiased estimators of poverty indices. Suppose that we have a sample of N households, to each of which is attached an expansion factor or weight, w_j. Given the correct choice of weights, an unbiased estimator of household poverty, for example, can be written

$$\pi = \frac{t}{p} = \frac{\sum_{j=1}^{N} w_j \pi_j}{\sum_{j=1}^{N} w_j}. \tag{1}$$

(For individual-level poverty measures, replace w_j by $w_j h_j$ in the denominator.) If the sample is self-weighting, then w_j is constant for all households and cancels.

3. Standard Errors for Poverty Measures

In contrast to the calculation of poverty index estimates, to make any progress estimating sampling variances one must be explicit about the type of sample design. Sample designs can deviate in many ways from the classical model of the simple random sample. We set out below a set of assumptions which both approximate many household survey designs and yield tractable standard error formulae. (Deviations from these assumptions are discussed at the close of the section.)

We allow for any number of explicit strata, with one stratum equivalent to no stratification, and further assume that the survey being analyzed:

(i) is clustered, with selection of clusters by probability proportional to estimated size (ppes)—with probability proportional to size (pps) as a special case;

(ii) uses two-stage sampling selecting first clusters (e.g. villages or street blocks) and then households drawn at the second and final stage as the ultimate sampling units;

(iii) was selected by random sampling with replacement (like blindly drawing a number from a hat and replacing it before the next draw) at the first stage;

(iv) was selected by systematic (every nth from a list) or random sampling at the second stage, with or without replacement, but with equal probability of selection; and,

(v) is self-weighting or has expansion factors which vary, at most, at the cluster level, due to the use of ppes and/or disproportionate stratification.

For a discussion of these features, see Howes and Lanjouw (1997).

Suppose the sample design divides the population into H strata ($H \geq 1$). In the hth stratum, let t_h be an unbiased estimator of total poverty in that stratum. Then

$$t = \sum_{h=1}^{H} t_h. \tag{2}$$

Let there be N_h clusters in a given stratum from which n_h cluster samples are chosen with replacement. Let M_{hc} be the true size of (number of households in) the cluster in the hth stratum from which the cth cluster sample is taken (the cth cluster for short). Let Z_{hc} be the estimated size of the cth cluster. The total estimated stratum population size is $Z_h = \Sigma_c Z_{hc}$. Randomly select one cluster with probability proportional to estimated size, i.e. with probability Z_{hc}/Z_h, and sample m_{hc} of the M_{hc} households. Define the cluster-level expansion factor, which is constant for all households, i, in cluster hc, as:[1]

$$w_{hci} = w_{hc} = \frac{Z_h}{Z_{hc}} \frac{M_{hc}}{m_{hc}}. \tag{3}$$

[1] $\Sigma_h \Sigma_c \Sigma_i = \Sigma_j$ and $w_{hci} = w_j$, etc.

Note that w_{hc} takes into account any or both of disproportionate stratification and the use of ppes. (If pps is used, $Z_{hc}=M_{hc}$ and, assuming m_{hc} to be constant within a stratum, the weights vary only between strata). Since the selection of clusters is random, any one cluster can be used to form an unbiased estimator, t_{hc}, of total poverty in the hth stratum

$$(4) \qquad t_{hc}=w_{hc}\sum_{i=1}^{m_{hc}}\pi_{hci}.$$

Note that this is the mean of the π_{hci} in the cth cluster times $Z_h*(M_{hc}/Z_{hc})$, the estimate of the size of the hth stratum based on cluster c. A more efficient estimator can be obtained by utilizing information from all cluster samples. From n_h cluster samples we can construct a combined unbiased estimator of total poverty in the hth stratum as the unweighted mean of the t_{hc}:

$$(5) \qquad t_h=\frac{1}{n_h}\sum_{c=1}^{n_h}t_{hc}.$$

The population size estimator, p, is defined analogously to t. Replace π_{hci} in (4) by either ones (if one wants a household poverty measure) or by h_{hci} (if one is after a per capita poverty measure). Then p_{hc} is exactly analogous to t_{hc} and p_h to t_h.

These definitions can be combined with a Taylor's expansion to provide a consistent estimator of the variance of π (see Howes and Lanjouw, 1997, for details on the derivation). We know that a Taylor's expansion and the definition of $\pi=t/p$ gives:

$$(6) \qquad \hat{\text{Var}}\,(\pi)=\frac{1}{p^2}[\hat{\text{Var}}\,(t)+\pi^2\,\hat{\text{Var}}\,(p)-2\pi\,\hat{\text{Cov}}\,(t,p)].$$

In the simplest case when one has srs and is measuring household poverty, so $p=N$, the number of households, $\hat{\text{Var}}\,(t)$ is the estimated element variance and (6) simplifies to the standard formula:

$$(7) \qquad \hat{\text{Var}}\,(\pi)=\frac{1}{N^2}\hat{\text{Var}}\,(t)=\frac{1}{N}\sum_{j=1}^{N}(\pi_j-\pi)^2.$$

However, for the more complex and realistic sample design considered here, Howes and Lanjouw (1997) show,

$$(8) \qquad \hat{\text{Var}}\,(t)=\sum_{h=1}^{H}\hat{\text{Var}}\,(t_h)=\sum_{h=1}^{H}\frac{1}{n_h(n_h-1)}\sum_{c=1}^{n_h}(t_{hc}-t_h)^2,$$

$\hat{\text{Var}}\,(p)$ is defined analogously, and

$$(9) \qquad \hat{\text{Cov}}\,(t,p)=\sum_{h=1}^{H}\hat{\text{Cov}}\,(t_h,p_h)=\sum_{h=1}^{H}\frac{1}{n_h(n_h-1)}\sum_{c=1}^{n_h}(t_{hc}-t_h)(p_{hc}-p_h).$$

This result, although little utilized in the poverty literature, has long been available with respect to sample means.[2] It is remarkable for its simplicity and useful for its very weak assumptions about sampling at the sub-cluster level. In particular, the variance formula takes no explicit account of sample variability within each cluster. The cluster estimates may be obtained using random sampling, with or without replacement, or systematic sampling. These formulae may also be used to calculate standard errors for poverty estimates for sub-populations of interest.

We now briefly consider relaxing some of the initial assumptions to cover a wider range of sampling designs. Clustering is almost universal, but there may be more than one level of clusters (that is, more than two stages of sampling). This poses no problem as long as the first-level clusters are selected with replacement. One takes the number of clusters above as given by the number of these first-level clusters. In this case, expansion factors may not be constant within the clusters thus defined, since lower-level clusters may be chosen using ppes, but unbiased estimates of total stratum poverty from each first-level cluster can still be obtained. Likewise if individuals or groups of households are the ultimate sampling units. Hence assumption (ii) is easily relaxed.

What if the absence of self-weighting is due to non-response (relaxing assumption (v))? If, as sometimes happens, non-response is corrected at the cluster level using weights, this is exactly analogous to having cluster weights due to ppes.[3] However in other cases, non-response is assumed to be a function of other-than-cluster-level variables, such as age groups. One is then left with weights which can vary within clusters. For discussion of this more difficult case see Hansen *et al.*, 1953; Kish, 1965; and Pudney and Sutherland, 1994. The arguments of the latter authors would suggest there is little loss in accuracy from treating these as random weights and thus in the same way as expansion factors which vary within first-level clusters due to the use of multi-stage clustering (see the previous paragraph).

We now come to the only restrictive assumption, (iii), namely that the sampling of clusters is random and with replacement. In fact, we know that many if not all surveys use systematic sampling at the first stage, including our two examples which follow. Why not assume this then? If the ordering on which the systematic sampling is based is approximately random, then systematic sampling will approximate random samples. However, what if sampling is systematic with non-random ordering—the case of "implicit stratification".[4] Since the randomness

[2]Kish (1965) is the classic on this subject. It gives what is still probably the most comprehensive treatment, though not the simplest. Som (1973) provides what we found to be the clearest presentation of results. Hansen, Hurwitz and Madow (1953) provide proofs. Levy and Lemeshow (1991) provide an introduction, as does Scheaffer, Mendenhall and Ott (1990).

[3]This is assuming that the respondents have the same characteristics as those questioned. In many contexts this will not be the case and non-response will introduce sample selection biases in estimates. There is no general solution to this problem. Note that to correct for non-response in this way, one needs to know whether existing expansion factors assume target or realized cluster sample sizes.

[4]Implicit stratification is called "implicit" to distinguish it from the "explicit" variety which is typically referred to (as in this paper) simply as stratification, and which is a completely non-random division of the population into different groups prior to any sampling (e.g. different provinces or income classes).

enters only once (at the start) of a systematic draw, unless one has several systematic draws (known as "replicated sub-sampling"), one does not have enough variation in the sample to estimate variances.[5] Since implicit stratification typically adds precision, for the same reason as explicit stratification, variances calculated under the assumption that the sample is random will tend to be upper bounds on the true variances and so will be conservative.[6] However, we can do better than this. If we know the original ordering of clusters, we can treat the samples *as if* they had been derived using explicit stratification using the following method given by Kish (1965), and detailed below.

Say that all the clusters in the hth explicit stratum have been listed from 1 to N_h and n_h have been selected systematically. Group pairs of clusters closest to each other into sub-strata (starting with the first two). Calculate $\hat{\text{Var}}\,(t_h)$ as the sum of estimated variances for each sub-stratum, where the latter are calculated using the pairs of clusters and equation (8). $\hat{\text{Var}}\,(t_h)$, thus calculated, would be an unbiased estimator of the hth stratum variance if that hth stratum had been divided into $n_h/2$ sub-strata—each containing, among others, two neighbouring sampled clusters—and two clusters had been selected at random and with replacement from each of these sub-strata. Moreover, $\hat{\text{Var}}\,(t_h)$ will also be only slightly biased upwards as an estimator of $\text{Var}\,(t_h)$ given actual sample design. There is no overestimation at all if the ordering of clusters within each sub-stratum is random. Since clusters which are close to each other in a list used for implicit stratification are unlikely to differ much, this assumption is unlikely to be seriously violated and so the overestimation will be small.

4. Examples from Pakistan and Ghana

What we are interested in measuring in this section are *sample design effects.* These give the ratio of the standard error one obtains given a particular set of assumptions about sample design to the standard error one obtains assuming that the sample is a simple random one. This ratio thus indicates the magnitude of the error made when ignoring sample design. The design effect also indicates the efficiency gain or loss from moving away from simple random sampling.[7]

[5]There are some samples which are based on replicated sub-samples. However, most surveys either are not designed in this way or, if they are, do not record which cluster is part of which replicate. Replicates can also be created by randomly sub-sampling from the sample once it has been collected. However, this must be done based on the design of the original sample. Taking random sub-samples which ignore any clustering or stratification present and basing variances on the variation in the mean estimates from these different sub-samples will lead to biased variance estimates. The method of "balanced repeated replications" (Kish and Frankel, 1970) is one type of sub-sampling which does take into account sample design. However, in this case one might just as well, wherever possible, use the same assumptions used to replicate (post survey) for calculating the variances analytically. It is only when this is not possible (say when one has statistics more complex than unconditional means) that post-survey sub-sampling is appropriate.

[6]The qualifier "tend" is required since one can think of cases in which implicit stratification will reduce precision. This will be the case if the elements display periodicity: so that every sth element will be similar or if the ordering displays very strong trends (Kish, 1965, pp. 120–21). Neither case will be relevant in the household survey context.

[7]A precise comparison requires sample data for the same population under the alternative sample design. The lack of multiple different surveys may bias the comparison. For example, the standard estimator for the element variance is biased downward in a clustered sample. However, one can show that such biases are negligible for the cases considered here (Kish, 1965; Deaton, 1997).

As noted in the introduction, ignoring stratification will lead to an *over*estimate of standard errors. On the other hand, ignoring clustering will lead to an *under*estimate of Var (π) if, as is typically the case, there is homogeneity within the cluster with respect to the variable of interest. For a given sample size, the variance of estimators is smaller the greater the homogeneity between clusters within strata and the greater the heterogeneity within clusters. The introduction of clustering increases within-stratum homogeneity, since clusters within a given sample differ by less than households do. This means that stratification becomes a more potent tool for increasing estimator precision once one has clustering (Kish, 1965, p. 164).

While the biases from ignoring stratification and clustering run in different directions, researchers have found that ignoring sample design typically results in standard error estimates which are too small (see, for example, Kish and Frankel, 1970, p. 1075). We turn now to two examples to get an idea of the magnitude of this bias with respect to poverty estimates.

The two surveys from Pakistan and Ghana are used because of the contrast they offer. Both are Living Standards Measurement Surveys (LSMSs), sponsored by the World Bank. Both are clustered, but the two surveys take different approaches to stratification. The Pakistani sample is designed in a traditional, highly-stratified way (described in Howes and Zaidi, 1994). Its 4,745 observations are collected from 300 clusters in 104 strata (see Table 1). The Pakistani sample is also non-self-weighting. Its cluster-level weights reflect the usage of ppes as well as disproportionate stratification, with, in particular, urban households over-represented.

The Ghanaian sample is made up of 3,181 households from 170 clusters. It is self-weighting on account of the use of varying cluster-takes based on the ratio of actual to estimated size (see Scott and Amenuvegbe, 1989, for a detailed description). It is not explicitly stratified at all. However, implicit stratification (i.e. systematic sampling with non-random ordering) was used: clusters were selected on the basis of a geographical ordering. We approximate the impact of this using the method recommended in Section 3. On this basis, the 170 clusters were ordered in the way they were in the sampling frame and divided into 85 strata, starting with the first two clusters in the first stratum.

In Table 2 we present standard errors for the means of household size and aggregate expenditure and the three most used poverty measures: the head-count, poverty gap and FGT2 indices. Both the poverty indices and the expenditure means are per capita. Since the exercise is purely illustrative we do not worry about equivalence scales or regional deflators, and arbitratory poverty lines are

TABLE 1

SAMPLE DESIGN FOR PAKISTAN AND GHANA LSMSs

	Explicit strata?	Number of clusters	Cluster weights?	Sample size	Average cluster sample size
Pakistan	104	300	Yes	4,745	15.82
Ghana	No	170	No	3,181	18.71

TABLE 2

SAMPLE DESIGN EFFECTS FOR MEAN EXPENDITURE, HOUSEHOLD SIZE AND VARIOUS POVERTY MEASURES FOR TWO SURVEYS

		Mean Household Size		Mean Expenditure		Head-count		Poverty-gap * 10		FGT ($\alpha = 2$) Index * 10	
		Pak.	Ghana	Pak.	Ghana	Pak.	Ghana	Pak.	Ghana	Pak.	Ghana
Estimate		7.17	4.52	5,934	71,276	0.336	0.333	0.920	0.980	0.364	0.403
Standard errors	A. Clustering and stratification	0.072	0.081	125	1,935	0.012	0.016	0.046	0.072	0.024	0.039
	B. Clustering w/o stratification	0.087	0.098	158	2,323	0.017	0.020	0.063	0.083	0.032	0.043
	C. Stratification w/o clustering	0.059	0.053	86	918	0.009	0.010	0.033	0.042	0.018	0.024
	D. Simple random sample	0.061	0.055	91	1,002	0.009	0.011	0.035	0.044	0.019	0.025
Ratio of standard errors	A/D	1.18	1.47	1.37	1.93	1.33	1.45	1.31	1.64	1.26	1.56
	B/D	1.43	1.78	1.74	2.32	1.89	1.82	1.80	1.89	1.68	1.72
	C/D	0.97	0.96	0.95	0.92	1.00	0.91	0.94	0.95	0.95	0.96
	A/B	0.83	0.83	0.79	0.83	0.71	0.80	0.73	0.87	0.75	0.91

Notes: See the text for explanation of the different standard errors.

set at 3,780 rupees per capita per month for Pakistan and 37,900 cedis per capita per annum for Ghana. Both lines put approximately 1/3 of the sample in poverty.

For each measure, four standard errors are given. Each is an estimate of the standard error assuming that the sample design is the one indicated on the left. The first ("stratification and clustering") is correct under the assumptions outlined at the start of Section 3 and the degree of stratification and clustering found in each survey (with the necessary approximations made where necessary to fit the surveys to this model). The second ("clustering without stratification") uses the same set of assumptions but ignores stratification. The third ("stratification without clustering") makes the same set of assumptions but ignores clustering and instead assumes a random sampling of households (with replacement) within each stratum. The fourth ("simple random sample") assumes that the surveys are simple random samples (with replacement). In all four cases, it is assumed that the survey is analyzed using the same set of weights, so actual poverty estimates are unchanged.

We begin with Pakistan. For mean household size, the design effect is not that large. Standard errors increase by only 18 percent over a random sample (A/D). This is what one would expect as there will typically be a great deal of variation within any one cluster with respect to household size. However, for mean expenditure and the poverty measures, the design effect is much larger. For the mean, it is 37 percent; for the head-count, 33 percent; for the poverty gap, 31 percent; and for the FGT2 index, 26 percent. On average, standard errors increase by around a third. Ignoring the sample design of the Pakistan survey leads to calculated standard errors which are substantially smaller than the correctly estimated standard errors.

It is clear from the table that stratification is very important. With clustered but unstratified data, standard errors would be between 68 percent and 89 percent (depending on the measure) greater than under srs (B/D). Note, however, that stratification is only important when the sample is clustered. Take mean expenditure as an example. Without clustering, stratification reduces the standard error on mean expenditure by only 5 percent from 91 to 86 rupees (C/D). With clustering, stratification reduces standard errors by four times as much, that is, by 21 percent from 158 to 125 rupees (A/B). Taking into account clustering but not stratification can be as misleading as taking neither into account. For the mean, the former strategy's estimate of 158 overestimates the correctly calculated standard error of 125 by as much as the latter strategy's estimate of 91 underestimates it.[8]

The results from Ghana are qualitatively very similar, and quantitatively larger. Sample design increases the estimated standard error of household size by 47 percent (A/D). For the other indicators, increases range from a low of 45 percent for the head-count to 55–65 percent for the poverty gap and FGT2 index to 93 percent for mean expenditure. Again the importance of taking stratification (here implicit) into account is evident. With clustering, stratification reduces estimated standard errors by 9 to 20 percent (A/B).

[8]Of course, one might prefer the former on the grounds that it is better to err on the side of caution.

The interest in constructing poverty measures is to make comparisons over time or space and it is here that the inaccuracies in standard error computation are of particular importance. Take the case of Pakistan. Say we had another survey of identical size and design a few years later. Assuming that the new survey would give approximately the same standard errors, how much of a change in poverty would the new survey have to register for that change to be judged significant (at a level of 0.05), that is, not due to sampling variation? For the head-count, for a simple random sample, the, say, decrease would have to be from 33.6 percent to 31.1 percent, a decline of 7.4 percent. With clusters and strata, a 10.1 percent decline would be required. One can imagine many cases in which two survey points would register a change in poverty of 7–10 percent: at least in Pakistan, these are the cases in which taking account of sampling design is imperative.[9]

5. Conclusion

Household surveys are far from being simple random samples. Typically, they are clustered and highly stratified, to name only the two most important departures from the srs model. Although recently published papers on statistical poverty analysis have ignored these departures, they are of considerable importance. The examples of the previous section show that taking into account sample design can increase standard errors by around one-half.

The main recommendation of this paper is therefore that standard errors for poverty measures should be calculated taking proper account of sample design. Both stratification (explicit or implicit) and clustering should be factored in. Taking account of only one, and not the other, can lead to inaccuracies as great as taking account of neither. Often the complexity of sample design will necessitate approximations, but these are much less egregious than just assuming simple random sampling.

The results also highlight the importance of making available information about sample design with the sample information itself. For example, households should have stratum and cluster identifiers. Expansion factors should be made available and their derivation documented. If implicit stratification is used to pick clusters, the relationship of the cluster numbering to the original ordering of clusters should be given. For many surveys, precisely because attention has not been paid to sample design by end-users, these basic requirements are not fulfilled.

References

Atkinson, A. B., On the Measurement of Poverty, *Econometrica*, Vol. 55, 1987.
Deaton, A., *The Analysis of Household Surveys: Microeconometric Analysis for Development Policy*, World Bank, 1997.

[9]Rodgers and Rodgers (1992) calculate sample design effects for permanent (multi-year) poverty measures using U.S. data and the balanced replication method mentioned in footnote 5 above. They find larger sample design effects of 50 to 200 percent (see their footnote 22). Comparing their results with those of this paper suggests the data source used by Rodgers and Rodgers is not highly stratified (so that the clustering effect is hardly offset at all), though the actual degree of stratification is not given in the paper.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory, Volume II—Theory*, John Wiley and Sons, New York, 1953.

Howes, S., Income Distribution: Measurement, Transition and Analysis of Urban China, 1981–1990, Ph.D. Thesis, London School of Economics, 1993.

Howes, S. and J. Lanjouw, Making Poverty Comparisons Taking into Account Survey Design: How and Why, Living Standards Measurement Study Working Paper no. 129, World Bank, 1997.

Howes, S. and S. Zaidi, Notes on Some Household Surveys from Pakistan in the Eighties and Nineties, Mimeo, The World Bank, 1994.

Kakwani, N., Statistical Inference in the Measurement of Poverty, *Review of Economics and Statistics*, 632–39, 1993.

Levy, P. and S. Lemeshew, *Sampling of Populations: methods and applications*, John Wiley and Sons, New York, 1991.

Kish, L., *Survey Sampling*, John Wiley and Sons, New York, 1965.

Kish, L. and M. Frankel, Balanced Repeated Replications for Standard Errors, *Journal of the American Statistical Association*, 65(331), 1071–93, 1970.

Pudney, S. and H. Sutherland, How Reliable are Microsimulation Estimates? An Investigation of the Role of Sampling Error in a UK Tax-benefit Model, *Journal of Public Economics*, 1994.

Ravallion, M. *Poverty Comparisons*, Harwood Academic Publishers, 1994.

Rodgers, J. and J. Rodgers, Chronic Poverty in the United States, *The Journal of Human Resources*, 28, 26–54, 1992.

Scheaffer, R., W. Mendenhall, and L. Ott, *Elementary Survey Sampling, Fourth Edition*, Duxbury Press, California, 1990.

Scott, C. and B. Amenuvegbe, Sample Designs for Living Standards Surveys in Ghana and Mauritania, LSMS Working Paper No. 49, World Bank, 1989.

Som, R. K., *A Manual of Sampling Techniques*, Heinemann, London, 1973.