NATIONAL LIVING STADARDS SURVEY
REVISED SAMPLE DESIGN: TECHNICAL DESCRIPTION

Peter Ellis and Chris Scott

March 1993

The World Bank
Southern Africa Department
Population and Human Resources Division
The Project for Statistics on Living Standards and Development

NATIONAL LIVING STANDARDS SURVEY

Revised sample design: technical description

1. Background

A preliminary document on the sample design was prepared in December 1992[1]. The present paper goes into fuller detail and is almost final. One or two details remain to be completed which will be available only when all the sampling frames are complete.

2. Theory

We begin by assuming the simplest case: a two-stage self-weighting design without stratification. The 1st stage units are census ESDs, the 2nd stage are households. Subscript **i** indexes ESDs, subcript **j** indexes households.

We use systematic sampling throughout - that is, sampling at a fixed interval in a list of units, starting from a randomly determined starting point.

Let N_i be the census population of the i-th ESD and let **N** be the total census population (all ESDs). Thus $N = \Sigma N_i$.

Let **a** be the number of ESDs to be selected and let b_i be the number of households to be selected in the i-th ESD.

We propose to select ESDs with probability proportional to size, that is to N_i. (See below for the method of doing this.) The probability that the i-th ESD falls in the sample will then be

$$p_{1i} = a\, N_i/N \qquad 1$$

The subscript 1 has been used here to indicate the 1st stage of sampling.

In each selected ESD we carry out a listing of households by means of a field operation. Let M_i be the number of

---

[1] "National Poverty Survey: Proposed sample design: technical description".

households listed in the i-th ESD

From those listed, we select a sample of b_i households by systematic sampling with equal probability. Thus the (conditional) probability of selecting household j in the i-th ESD will be

$$p_{2i} = b_i / M_i \qquad \text{---------- (2)}$$

(Subscript j does not appear because all households within this ESD are selected with the same probability.)

The overall probability of selecting household j in the i-th ESD is the product of these two probabilities, namely

$$p_{1i}\, p_{2i} = a b_i N_i / (M_i N)$$

Since the sample is to be self-weighting this expression must be equal to a constant which is the overall sampling fraction **F**. Thus:

$$F = a\, b_i\, N_i / (M_i N) \qquad \text{---------- (4)}$$

The overall sampling fraction must be the ratio of the number of households in the sample to the number in the population. The number in the sample has been fixed by the survey organizers; let this be **m** . The number in the population has to be estimated. Let it be **M**. Theoretically $M = \Sigma M_i$, where the summation is taken over the whole population, but the value of this is unknown. Thus **M** has to be estimated from other sources. If our estimate of **M** is incorrect we will fail to get the desired sample size, but there will not be a bias.

Thus we have:

$$m / M = a\, b_i\, N_i / (M_i N) \qquad \text{---------- (5)}$$

3. Procedure for selection of area units with probability proportional to size (PPS)

1. List the area units with the size against each size census population = N_i).

2. Cumulate the N_i variable in a column on the right. Check that the last cumulated value entered is equal to the total **N**.

3. Compute the 1st stage sampling interval I_1 needed to yield the desired number **a** of area units, using $I_1 = N/a$

4. Select a random number between **1** and $\mathbf{I_1}$: let this be **R**. Compute the sampling sequence:

$$R \; ; \quad R + I_1 \; ; \quad R + 2I_1 \quad R + 3I_1 \; ;$$

5. Use this sequence to select the sample, thus: for each term of the sequence find the first number in the cumulative column which equals or exceeds that term. The unit corresponding to this number is the one selected.

4. Procedure for selecting households with equal probability

Select at a fixed interval in the household list, starting from a random number less than or equal to the interval. For computation of the interval $\mathbf{I_2}$, see Section 5. In many cases the interval will be small - typically under 5. Rounding such intervals to the nearest whole number would involve an unacceptable degree of error. It is proposed therefore to compute the interval correct to 1 decimal. Selection at a decimal interval may be done as follows:

1. Number the units serially: 1, 2, 3

2. Let $\mathbf{I_2}$ be the interval, to the nearest 0,1.

3. Multiply $\mathbf{I_2}$ by 10: select a random number between 01 and 10 I_2. Insert a decimal comma before the last digit of this. Let the result be **R**.

4. Write down the sequence of sampling numbers:

R
$R + I_2$
$R + 2I_2$
$R + 3I_2$
etc.

5. Ignoring the decimals in the sequence, the whole number part of each sampling number indicates the unit selected.

Example. Let the interval be $I_2 = 3,4$
Select a random number between 01 and 34, say 23
Then R = 2,3
The sequence is 2,3
5,7
9,1
12,5
15,9
etc.

Read off the whole numbers to give the units selected: 2, 5, 9, 12, 15 ... It will be seen that the method gives an interval which is sometimes 3, sometimes 4, with the desired average of 3,4. Note: if the random start R is less than 1 the first term of the sequence yields no selection; the method is still valid, however, with R + I giving the first selection.

5. Parameters

In practice the area units will not always be ESDs, as will be made clear later. Sometimes they will be blocks delineated on maps, in other cases villages or village groups. For the sake of generality, we refer henceforth to the last-stage area units as the "Ultimate Area Units", or UAUs. The majority will in fact be ESDs.

Total sample size **m** has been fixed at 9000 households.

The total census population (de facto, adjusted, in households and migrant hostels, including TBVC) comes to 38 120 853. The current population total $\mathbf{N_{93}}$ is estimated from the adjusted census total (with separately estimated totals for the TBVC states) extrapolated at a rate of increase of 2.5% per year over a period of 2.3 years. This comes to 40 350 000. Dividing this by the mean national household size **H** will give an estimate of **M**. The value of **H** is particularly uncertain: our best estimate is 5,00, which gives 8 070 000 for **M**.

The mean cluster take, that is the average number of households to be selected per UAU, has been fixed at 25.[2] Dividing this into the sample size of 9000 yields a total of 360 sample UAUs. This is therefore the value of **a**.

The sampling interval $\mathbf{I_1}$ for selection of ESDs is obtained by dividing the total population in the frame, namely **N**, by the number of units it is desired to take into the sample, namely **a**. This yields 105 800.

The sampling fraction at the household stage is $\mathbf{b_1/M_1}$. More useful operationally is the household stage sampling interval, $\mathbf{I_{21}}$, which is the reciprocal of the sampling fraction, or $\mathbf{M_1/b_1}$. From formula (5),

---

[2] For the reasoning behind this choice, see Chapter 3 "Sample Design", pp.24-25, in the World Bank SDA Working Paper No.14: The Social Dimensions of Adjustment Integrated Survey, 1992, Washington DC.

$$I_{2i} = M_i/b_i = aMN_i/(mN)$$

Substituting $m/a = 25$ and $MH = N_{93}$ we obtain

$$I_{2i} = (N_{93}/N) \cdot (1/25H) \cdot N_i$$

Finally, inserting the numerical values mentioned above,

$$I_{2i} = N_i / 118{,}1$$

We recall that N_i is the (adjusted) 1991 census population figure for the i-th UAU. In the majority of cases the UAU is a census ESD; for the procedure where it is not, see sections 7 and 8. Pending that discussion, we see that formula (8) enables us to compute the 2nd stage sampling interval I_{2i} for each selected ESD before any field work begins. This interval does not depend on the number of households listed in the i-th ESD. Thus, we can give the interval I_{2i} to the field team, together with a random number between 1 and I_{2i}, and the team supervisor can make the household selection as soon as the household list in the ESD is ready without the need for further communication with headquarters.

Finally, note that the method does not assume the correctness of the census, because the listing operation provides an updating of the ESD population (in terms of current households).

6. ESDs that are too small

The method described above assumes that 25 households can be selected in each UAU. More exactly, looking at formula (8) we see that if N_i ~~exceeds~~ 118 we are going to have a sampling interval for households which is less than 1. Such an interval cannot be achieved unless we are willing to introduce weights; but since a self-weighting sample has been predicated we should avoid this situation.

It can most easily be avoided by combining any ESD having a census population of 118 or less with another, preferably neighbouring, ESD. Since the ESDs in the sampling frame are listed by Magisterial district (MD), it is proposed to group any such small ESD with the next ESD on the list. In most cases the two will not be contiguous but will not be far away, since they are in the same MD. (Exception: if the small ESD is the last one listed in the MD, it should be combined with the preceding one, not the next one.) From the method of selection described in Section 3 above, it is clear that we do not need to identify such cases except where a selection is made: in the regions of the list between selections, the

combining of ESDs will have no effect. Thus there is no need to go through the whole sampling frame looking for ESDs with a population of 118 or less: it is sufficient to do the selection first and then see whether any such ESDs have been selected. If so, they must be combined with their neighbour in the list.

However, this procedure still leaves one loose end. Suppose ESD number i has a population under 119. If it is selected it will be grouped with ESD i+1, which may be of normal size. But ESD i+1 already has a chance of being selected on its own account. Thus it has two chances: one on its own account, one through ESD i. The selection probability for ESD i+1 is therefore not N_{i+1} but $N_i + N_{i+1}$. Thus the check on small units has to be done in both directions: for every ESD selected, even if it is large, we have to check to see whether it is preceded in the list by a small ESD which would have been combined with it had that small ESD been selected, and if so, the two have to be combined. As there are only 360 units to be selected in all, this is not an undue amount of work.

Thus, when two ESDs have to be combined, their two size measures N_i and N_{i+1} must be added together before applying formula (8) to compute the household sampling interval. Also the two household listings have to be combined before the household sampling is done. As far as the survey records are concerned it is best that they be treated as a single area unit: it is only for the purpose of field work organization that we need to retain the fact that two areas are involved.

. Case of 3-stage sampling

In some areas 2 stages of area sampling may be necessary, making 3 stages in all.

The above schema can be very simply modified to deal with this situation, provided that an estimate $\mathbf{N}_{1i}$ is available for each 1st stage unit (usually termed primary sampling unit, or PSU). Assuming this, it should always be possible to construct a set of 2nd stage area units (SSUs) in each PSU, using maps or field visits. Once this has been done, it will always be possible to make an estimate $\mathbf{N}_{2ij}$ of the population of each SSU (where j now indexes the SSU), if necessary by assuming that all the SSUs in the selected PSU are equal. Procedures will be simplified if we adjust the estimates $\mathbf{N}_{2ij}$ within the selected i-th PSU so that they sum to $\mathbf{N}_{1i}$, ensuring that the estimates at the 1st two levels are consistent. If this is done we can simply replace each selected PSU i by the

list of SSUs ij which it contains, for each of which we have a size estimate. The same sampling sequence as before thus selects one of these SSUs without any modification of the sequence.

This method assumes that the survey planner wishes to select only 1 SSU in each PSU. The method should be applied for any selected area unit judged to be too large for listing. The limit for single stage area sampling (2 stages in all) might be set at 500 households. This means that any unit selected whose census population exceeds 2500 persons should be subdivided into SSUs, only one of which will be selected.

8. The case of separate sampling frames

The methods described above assume that we have a single area-sampling frame covering the whole country. However, in the present survey we will have to use at least 3, possibly as many as 6 separate frames since the census was not done in the same way everywhere. The first, and much the largest, frame will be the census; the others will be frames for ~~some of~~ the individual homelands. These various frames will be examined in the following sub-sections and the allocation of the sample between them will be computed. We begin by recalling that the sampling interval for area-stage selection is 105 800 census population. This implies that the number of UAUs to be selected in any area or grouping can be obtained by dividing its census population by 105 800.©

using five frames

in TBVC

Main census frame

This frame consists of a list of all the census ESDs, grouped according to MD, urban/rural, statistical region and province in that hierarchical ascending order. The list shows one line per ESD.

In this frame we select a single stage sample of ESDs with probability proportional to census population, in the standard manner described above.

Sub-frame within the census

Most of the black metropolitan areas were not covered exhaustively in the census. Instead, a sampling method was used. Aerial photographs were taken of the zones concerned and on the basis of these the zone was divided into blocks, of

---

© For a detailed breakdown of the number of UAUs selected from various areas, see Sampling Frame: Number of clusters per Region

approximate size 30 - 60 stands.[3] All blocks were delineated and numbered on transparent plastic overlays placed on the photos. The sizes of the blocks (number of stands) were estimated from the aerial photos and a number of stands were selected with probability proportional to size.

This method can be adapted to the needs of the present survey. Each of the 91 zones covered by this method appears as a single entry in the census list mentioned in sub-section 8.1, just as if the whole zone were one ESD. (Even Soweto appears as just one ESD.) When carrying out the sampling procedure for ESDs, as in sub-section 8.1 above, the process will from time to time select one of these zones, perhaps once or perhaps several times. The number of times the sample series falls in the zone indicates the number of blocks to be selected in that zone. We will communicate this number to the team at HSRC who will select that number of blocks in the zone from the blocks marked on the photo overlays, using the same method as that used in the census. If up-dated photos are available these should be substituted.

The blocks will still be selected with probability proportional to size, but the measure of size will be different from that used in the first sampling stage: it will be the number of stands rather than the census population. This change of measure in no way invalidates the sampling (since the number of area units to be selected in the zone has already been fixed), but it does affect the computation of the sampling interval for selection of households in each area. To adjust this computation we have to bring the two divergent measures of size into line. This can be done by multiplying the block measures of size (say N_{ij}^* for the j-th block in the i-th zone) by a constant inflation factor which makes their total, $N_i^* = \Sigma_j N_{ij}^*$, equal to the corresponding total N_i for the other measure. In other words, the adjustment factor equals:

$$N_i \; / \; N_i^*$$

With this adjustment we can slot in the block list so that it exactly replaces the entry for the whole zone in the list of ESDs. The total remains the same; we have simply replaced the single line representing a huge area by the full details: numerous lines each representing a block. The fact that the list changes from ESDs to blocks at this point makes no difference: we have nowhere assumed that the area unit is of the same nature throughout the list.

---

[3] A stand is a plot originally intended for one household. The majority of stands do contain just one household but many contain two, and in very crowded areas a stand may contain numerous households.

Introducing this more detailed sampling frame to replace one part of the list, we have to replace N_i by N_{ij}* and apply the adjustment factor, in formulae (7) and (8). Thus the new household sampling interval is:

$$I_{2ij} = aMN_{ij}^*/(mN) \times (N_i/N_i^*$$

$$aMN_i/(mN) \times (N_{ij}^*/N_i^*$$

$$N_i / 118,1 \times (N_{ij}^*/N_i^*)$$

8.3 Frames in the TBVC ~~states~~

The frame may differ in its characteristics in each state. In this sub-section we deal with the particular case of Transkei. The procedures for the other states are similar but vary in the details.

In Transkei the only available frame giving measures of size for small areas is the data set resulting from the census demarcation operation. This was carried out in October-November 1990, while the census itself was carried out in December 1991.

In the demarcation data set we have a list of villages classified by Administrative area (AA), with the AAs classified by Magisterial district (MD) and for all of these units we have the estimated number of households. This set is available, however, for the rural sector only (which is about 95% of the Transkei population). The census data themselves are also available down to MD level but not below. The demarcation (rural) data in terms of households fall short of the corresponding census data by 19% but they could be corrected to agree with the census by introducing an adjustment factor computed at the MD level. However this problem can be solved more elegantly by treating the MDs as a sampling stage and adding them to the main census. Unfortunately the available census data on MD populations are on a de jure basis while we require de facto: the difference is very substantial. A better source would be the MD numbers of households from the census, which are available. We also have available an estimate of the total de facto census population, urban and rural separately, for the whole of Transkei. We can use these population estimates first, to determine the number of UAUs to select in rural and urban Transkei respectively, then move to the MD households to determine the number falling in each MD, and finally move down to the demarcation data to select villages (rural sector) or towns/suburbs (urban sector). Details follow.

1. Census population (de facto, excluding institutional):

Rural: 3 292 602
Urban: 182 534

2. Dividing by the sampling interval I_1 = 105 800 used in the main sample, this gives us the required number of UAUs to be selected:

Rural: 31
Urban: 2

3. Census households:

Rural: 639 822
Urban: 50 226

4. Sampling interval I_2 (divide data in 3 by data in 2):

Rural: 20 639
Urban: 25 113

5. Use these intervals to select MDs with PPS. Each MD will be selected 0, 1 or 2 times. Carry these numbers down to the next stage.

6. Rural. Move to the demarcation sampling frame. Work on each MD separately. For each MD selected in step 5, obtain the total number of households. Divide this by 1 or 2 (the number of selections of that MD in step 5) to get the sampling interval I_3. Use this interval to start a sampling with PPS. Begin at the AA level. Cumulate the AA totals within the MD and find out which AA is selected. Within that AA, wherever a village is below 50 households in size combine it with the next village on the list, using brackets. This gives a list of villages and village groups (VGs). Write in the size for each VG. Complete the cumulative size series for these villages and VGs, adding on to the last cumulative size computed for the preceding AA. Thus find out which village or VG is selected.

7. Urban. Note that only 2 selections fall in Transkei urban. Thus the matter can be treated on an ad hoc basis, but for the sake of rigour we describe here the preferred procedure in detail. Three distinct cases might arise: 1) The urban area selected may be Umtata, Butterworth or Ezibeleni (urban part of MD Lady Frere). These three towns have been subdivided for the census. 2) The urban area selected may be another town greater than 400 households in the census. Any such town selected will need to be subdivided. 3) The area selected may be a town less than 400 in size. This does not need to be divided. We deal with these 3 cases in turn.

7.1 These three towns were subdivided into suburbs, whose

size was estimated in the demarcation in terms of the number of census enumerators needed. These measures of size can be used for PPS sampling. Each unit represents an estimated 80 households. Step 5 shows the number of units to be selected (0, 1 or 2). If 0 the town is not selected. If 1 or 2, divide this into the total of the size measures for the town, to give the sampling interval I_3. Use this for PPS sampling to select suburbs. If the selected suburb is of size 5 units or less no further subdivision is needed; if larger, we will approach the Transkei Central Statistical Office for assistance in subdividing the area yet again.

7.2 If a town greater than 400 census households is selected, we approach the Transkei CSS for assistance in subdividing the town, then procede as above.

7.3 If a town less than 400 census households is selected this is accepted as the UAU.

It remains to determine how to compute the household sampling interval.

The procedure is essentially the same as in sub-section 8.2: we apply formulae (7) or (8) with adjustments analogous to those in (7a) and (8a).

If S is the number of area sampling stages, the final-stage (= household) sampling interval I_{s+1} is given by:

$$I_{s+1} = (N_{93}/mH) \cdot p_1 p_2 p_3 \cdots p_s$$

For any sampling stage s the conditional selection probability (probability of unit ijk being included in the sample given that the stages prior to the s-th have been selected) for unit ijk is given by:

$$p_{s,ijk} = a_{ij} N'_{ijk} / \Sigma_k N'_{ijk}$$

where a_{ij} is the number of s-th stage units to be selected in the (s-1)th stage unit ij, N'_{ijk} is the "size" of the unit ijk (the prime is used to indicate that the measure of size is not necessarily the same as in the 1st stage, whose size measures are N_i), and the summation Σ_k is taken over all units ijk existing in unit ij. (In the case where s = 1 it will be seen that this gives formula (6), bearing in mind that $N_{93} = MH$.)

Using this methodology, the sampling interval for households (or for stands) must be computed at SALDRU for each UAU and communicated to the agencies responsible for selection of households (stands). Note that, whether households or stands are used as the sampling unit at the final stage, the

sampling interval remains the same, provided that, where stands are used, the sample includes all households in each selected stand. The decision to use one type of unit or the other can therefore be left to the field agency in every UAU

9. Stratification

Stratification means division of the sampling frame, prior to sampling, into zones or "strata" which are as homogeneous as possible. The sample is then selected independently in each stratum.

With equal probability sampling (also termed self-weighting sampling, or proportionate sampling) the effects of stratification will be modest. The role of stratification would be limited to ensuring that the number selected in each stratum corresponds exactly to the number planned. Without stratification the two would correspond approximately, but with some random variation. The gain in precision resulting from stratification is greater where the households differ (in terms of the survey variables) more between one stratum and another and less within each stratum. Obviously in RSA race would be an effective stratifying variable because of the wide differences in income and welfare between racial strata. It is therefore important to get the racial distribution in the sample closely aligned with the racial distribution in the population.

Given equal probability sampling, systematic selection produces a stratifying effect without the need for creating explicit strata. This can be achieved by ordering the units before selection in such a way that all households of each race come in the same part of the list, or that all ESDs of a given composition come together in the list. This effect is generally termed "implicit stratification".

We therefore propose an implicit stratification at the area stage and another at the household stage.

At the area stage, the ESDs will be arranged in order of decreasing "per-cent black", within any given region. At the household stage the households in a selected ESD will be grouped by race: Black, White, Coloured, Indian. This procedure aims to produce the closest possible fit between the sample and the population as regards racial composition.

10. Sampling the migrant hostel population

In the areas which were enumerated in the census only on a sampling basis, the HSRC has a complete list of hostels with estimates of their size in terms of the number of beds. The list is kept reasonably up to date. In all cases the hostel constitutes one or more separate blocks: that is, no block shows a mixed area of hostel and non-hostel populations.

Still within the 91 HSRC areas, we propose to select separate samples of hostel-blocks and non-hostel blocks. The former presents yet another sampling frame. The two frames should be slotted together in the same manner as suggested in Section 8. Hostels will be considered as groups of 1-person households. (Exception: a few have been converted into tenement flats, where families are allowed.) There will be a serious problem of access, both for listing and interviewing, but there is no problem in defining the sampling method assuming some kind of listing can be achieved.

In the areas covered by the full census there will be very few migrant hostels and perhaps none will be selected. If any are, they can be treated as a set of households all living at one address. There is no need to make separate arrangement for sampling. To cater for this we have added to the listing form a 5th "race": hostel dwellers. (See the last five lines of Section 9 above.)