Expenditure and Service Delivery Study for the Education Sector

A Note on the Sample Design for the School Survey

09 April 2002

Introduction

A central part of the Expenditure and Service Delivery Study for the Education Sector (ESDS-ES) will be a survey of 200 schools across Zambia. In parallel with this school survey and in the same period, the 2001 National Assessment Survey of Grade Five Level (NAS) will be repeated, providing a unique opportunity to link outcome data on school and pupil performance to information of school functioning and financing. Since this link between the NAS and ESDS-ES is crucial for the study, it will be optimal to start from the sampling frame of the NAS and adapt it to our purposes, rather than trying to match an independent sample design of the ESDS-ES to the data of the NAS. Details on the aims of the overall study and further motivation of the link between the ESDS-ES and the NAS can be found in the Aide Memoire for the ESDS. In this note, the sample strategy will be documented. First, the NAS sample design will be explained and then the strategy for the ESDS-ES sampling will be explained.

The sample design of the National Assessment Survey

The aim of the NAS was to give “accurate ... information at the national, province and district level” on school and pupil performance with grade five enrolment. To achieve this, the sample design implied stratification by province and district level, as well as by urban-rural locality. The overall sample size was fixed in terms of schools, at 400, based on the population of schools with grade five enrolment. In each school a number of pupils at grade five were tested and interviewed.

frame

The 1999 Annual School Census provided the list of all schools with grade five enrolment, 3933 schools in total. However, the actual frame was in terms of the actual grade five enrolment, a total of 200524 pupils across the nation in 1999.

stratification

Province and districts are well-defined administrative classification so stratification of the school population is self-evident. The NAS used the Central Statistical Office definition of rural and urban. Details are in Sinyenga (1991). The

frame was completed by adding stratification codes by province, district and rural-urban.

sample allocation over strata

The stratification variables provide the domains (strata) and reporting levels of the survey. The number of schools per domain was allocated using the 'optimal square root method'. This method provides a middle way between proportionate allocation (i.e. selecting schools in each domain proportional to the domain size in terms of enrolment relative to the total population of grade five enrolled children) and equal allocation (an equal number of schools in each domain), and providing a minimal domain sample size for the smallest domain. The formula used is:

$$n_h = \max\left[b, nk\sqrt{W_h^2 + H^{-2}} \right] \qquad (1)$$

with n_h the sample size in domain h, b is the minimum sample size, k is a constant, n is the total desired sample size, W_h is the relative domain size, H is the number of domains. For the sample allocation across provinces, b is 34 schools, n is 400 schools, W_h is the grade five enrolment in the specific province relative to the total grade five enrolment (i.e. 200,524), while H is the number of provinces (9). The proportionality constant k is for the allocation across provinces was then about 0.695. Within each province, the minimum sample size per district was 4 schools. Across rural and urban areas, proportional allocation within each stratum was used, starting from the province. Table 1 provides the sample allocation across provinces; Sinyenga (2001) has details for the allocation within provinces[1].

[1] Note that the actual sample allocation in table 1 is affected by practical considerations and the minimum sample size per district, so that the formula in (1) is *ex-post* not exactly correct. In particular, in Southern and Northern Province, the sample is smaller (by 3 and 8 schools) than implied by the formula and in Copperbelt and Lusaka it is larger (by 3 and 6 schools). The rest of the deviations are rounding errors.

Table 1 Sample allocation of the NAS to Provinces

provinces	number of schools	grade five enrolment	sample allocation
Copperbelt	333	38483	56
Central	439	23524	44
Lusaka	221	27482	46
Southern	559	29097	54
Luapula	346	13793	36
Northern	697	24900	54
Eastern	548	19437	40
N/Western	345	10296	34
Western	445	13512	36
Total	3933	200524	400

actual sampling

With the total numbers of schools to be included in the sample fixed per stratum, the task remained to select the actual schools to be included in the sample, and within each school the pupils to be interviewed and tested. The most practical solution – two-stage sampling - was used for this purpose: the first stage sampling units were the schools and the second stage units were the grade five pupils.

The schools were selected using probabilities proportional to the estimated size of grade five enrolment per school. Size weighted school selection implies that at this stage, a grade five pupil in each stratum has equal probability for its school to be selected.

During the second stage, pupils where randomly selected from each sample school, with the total sample size up to 20 grade five pupils. Where a school has less than 20 grade five pupils, all of them were be selected and included in the sample. Note that this procedure implies a nonzero but non-equal probability that a pupil in grade five in each stratum will be selected in the sample. Consequently, weights have to be calculated for reporting results at the strata and national level. Details are in Sinyenga (2001).

Note that depending on the variables of interest, weights will have to be adjusted. The sampling frame contains sufficient information to do this. For example, when the unit of analysis is the school or the district, different weights will have to be used (since the selection probability of a district or a school is different from the selection probability of a pupil).

Linking the ESDS-ES to the NAS

Cost and logistical reasons imply that the ESDS-ES detailed school questionnaires will only be implemented in 200 schools. Nevertheless, the information provided by the NAS will remain crucial so only schools covered by that survey will be included in the ESDS-ES. Indeed, studying other schools will result in information on key school quality outcome indicators – based on pupil test performance – to be missing.

The issue therefore is first to determine the rule to include schools from the NAS in the ESDS-ES. A few different scenarios are discussed and problems are highlighted in the next few paragraphs before presenting a more detailed proposal for the actual procedure. First, simply reducing the number of schools surveyed per stratum is not feasible for the lowest stratum (the district), since in the NAS, 34 out of 72 districts have reached the 'minimum' number of schools per district (4 schools). Reducing this further (e.g. halving) would result in large sampling errors per district and jeopardise the function of districts as reporting levels. Note further that an equal number of schools per district would imply less than 3 schools per district anyway, since there are 72 districts to be covered. Consequently, a reduction of the total number of districts appears necessary.

Of course, this will imply that districts are not going to be reporting levels (and strata) any more, but rather clusters in a multi-stage sampling procedure. So which districts should be chosen? A few alternatives present themselves. First, we could retain the province as a stratifying variable. Then, in each province, size weighted random selection could be used to select districts until the sum of all schools investigated in these districts in the NAS equals *half* the number of schools allocated to each province in the NAS (as reported in table 1). Note that this is equivalent to using equation (1) but reducing the desired sample size n to 200, but retaining the minimum b[2]. The sample allocation to each province using (1) is summarised in table 2.

[2] An alternative procedure, reducing b as well, e.g. to 2 per district, is plausible as well, but the gain in precision for province data via the within province stratification would come at a high cost and difficult logistics (the need to cover each district in the country).

Table 2
Possible Sample allocation of the ESDS-ES to Provinces, retaining Provinces are Strata (based on (1) and sample size of 200)

provinces	number of schools	grade five enrolment	"optimal" sample allocation
Copperbelt	333	38483	31
Central	439	23524	22
Lusaka	221	27482	25
Southern	559	29097	25
Luapula	346	13793	18
Northern	697	24900	23
Eastern	548	19437	21
N/Western	345	10296	17
Western	445	13512	18
Total	3933	200524	200

In all surveys, cost and logistical reasons constrain the construction of ideal sampling frames. An alternative strategy would be to give up on provincial level reporting levels and simply use provinces as another cluster level in multistage sampling. The loss would be in terms of sampling errors for national results, while the gain could come in terms of the ability to cover all districts in the remaining provinces (allowing within-province district comparisons). This is particularly useful, since one key interest of the study is to how differences in district level characteristics, in terms of the functioning and actions of the district educational authorities, affect schools and pupils. Another important gain would be in terms of transportation costs via the ability to restrict the sample to well-defined geographical areas.

Technically, this would mean: the primary sampling unit is the province, followed by the school and then the pupil. Strata would be the district and rural-urban locality. It is possible to introduce a further stratification variable, whether the province is part of the decentralisation process of fund disbursement and decision making to the district-level (currently implemented in Lusaka and the Copperbelt, and soon to be introduced to the Northern, Southern and Western Provinces). This would suggest that Copperbelt and Lusaka would be included. The selection of the other provinces will depend on the total province sample size considered. For logistical reasons, it is likely to be restricted to 4, including Copperbelt and Lusaka. While random sampling (using appropriate rules) is feasible, practical and implicit stratification reasons should dominate for a draw of two provinces from effectively a small population of 7. Below in table 3, we give some other characteristics that may influence our choice, based on the 1998 LCMS (Central Statistical Office, 1998).

Table 3 Selected Province Characteristics

	Net enrolment age 7-13 (%)	Poverty head count (%)	% rural	Population share (%)
Central	75	77	66	10
Copperbelt	76	65	23	18
Eastern	49	79	91	13
Luapala	61	81	86	7
Lusaka	79	53	19	15
Northern	60	82	84	12
North Western	66	76	86	6
Southern	73	75	80	12
Western	64	89	90	7
All Zambia	68	73	63	100

Copperbelt and Lusaka are the largest and most urbanised provinces in Zambia, with the highest net enrolment rates and lowest poverty rates. This would suggest that it is appropriate to select more rural and poorer provinces with lower enrolment rates to complete the sample. Based on this notion, Eastern and Northern (the next largest provinces as well as those with the lowest enrolment rates and high poverty) were chosen to be part of the sample, and the example below shows how this works.

Based on this selection, optimal square root allocation (see formula (1)) between these four provinces and across all districts of these provinces gives the allocation reported in column four. However, if the minimum district sample size is maintained at 4, Northern would require at least 7 more schools in its sample. Once this is taken into account, (see the fifth column), then it is striking the sample allocation is virtually identical to the sample allocation to these four provinces in the NAS[3]. Since the matching the sample design of the ESDS-ES as close as possible to the NAS is a prime objective, this would provide strong support to use their sample allocation for these four provinces based on a careful weighing of practical and statistical reasons, even if it implies a slight reduction of the total sample size to 196. For sub-province levels, the sample design of the NAS can then be fully used.

[3] Recall footnote 1, stating that the NAS design also deviated somewhat from the optimal formula.

Table 4 Sample allocation of the ESDS-ES to selected Provinces

provinces	number of schools	grade five enrolment	"optimal" sample allocation	corrected sample allocation	sample allocation in NAS
Copperbelt	333	38483	60	57	56
Lusaka	221	27482	50	48	46
Northern	697	24900	47	54	54
Eastern	548	19437	43	41	40
Total	1799	110302	200	200	196

The sampling strategy summarised in table 4 implies that the *only* additional correction that is needed for any descriptive statistics and statistical analysis will be to account for the change of provinces as strata to clusters when calculating *overall* ("national") results from the data. In both descriptive and multivariate analysis, the corrections to (respectively) standard errors and variance-covariance matrixes are straightforward although cumbersome, but effectively handled by standard statistical packages such as STATA.

For certain purposes, such as reporting results on school functioning and school-district relations, the use of the grade five enrolment as the relevant population may not be ideal. Total school (grade one to six) enrolment may be more appropriate. This would suggest that *ex-post* the sampling frame could be supplemented with total school enrolment information. Correction weights for the the sampling probabilities could be constructed for the analysis.

References

Central Statistical Office (1998), *Living Conditions in Zambia – 1998*, Lusaka: CSO, processed.

Sinyenga, G. (2001), "Sample Design Procedures for the 2001 National Assessment Survey of Grade Five Level in Zambia", Central Statistical Office

World Bank (2002), "Aide-Memoire- Expenditure and Service Delivery Study Education Sector- Lusaka, December 8th 2001- December 22nd 2001", processed.