Sample Design and Estimation Procedures for Poverty and Social Impact Analysis (PSIA) Follow-up Study[1]

David J. Megill
Sampling Consultant
World Bank

June 2009

1. Background

The Poverty and Social Impact Analysis (PSIA) Follow-up Study was designed as a longitudinal survey based on a subsample of households interviewed in the 2002/03 *Inquérito aos Agregados Familiares* (IAF), a national household income and expenditure survey conducted in all provinces of Mozambique from July 2002 to June 2003. The objectives of the PSIA survey include estimating the impact of lowering the costs of primary education and measuring trends in other educational patterns such as the school enrolment rates. More details regarding the objectives of this survey are described in the Terms of Reference for the study. The earlier baseline study was based on the educational data from the 2002/03 IAF.

The purpose of this report is to document the sample design and weighting procedures for the PSIA Survey. Given that this survey will be based on a subsample of the households selected for the 2002/03 IAF, a summary of the IAF sample design is presented here as well as details of the sampling strategy that was used to select the subsample of households for the PSIA follow-up study. The PSIA survey design was developed with the collaboration and support of Louise Fox, Melissa Sekkel and Rui Benfica of the World Bank.

2. Nature of a Longitudinal Survey

The main advantage of conducting a longitudinal survey for the PSIA follow-up study is that it makes it possible to link the survey data with the corresponding IAF microdata for the same sample households. The correlation in the data from the two surveys should provide more precise estimates of trends (differences) in the indicators between the two surveys. The use of an existing sample of households also reduces the cost of developing the sampling frame, since no new listing of households is required. Given that the domains of analysis will be limited to the national level, urban and rural, it is only necessary to conduct the longitudinal survey in a subsample of the segments selected for the IAF.

At the same time, a longitudinal survey presents additional challenges for the data collection and estimation procedures. Given the mobility of part of the population and

[1] The sample and study was later renamed to Mozambique Education outcomes National Panel Survey

changes in household composition over time, there may be a substantial proportion of the original sample of households that are difficult to find or no longer exist in the same location, and there may also be some refusals. It is difficult to predict the corresponding attrition rate, but based on the experience in other countries, it could be as high as 15 to 20 percent. This loss in the effective sample size should be taken into account in the sample design. The attrition rate is generally higher in urban areas, given the greater mobility of the urban population.

For the analysis of the longitudinal data, it is also important to understand the level of inferences that can be made from the survey estimates. The sample for the longitudinal sample will only represent the more stable households that remain in the same location over the period of approximately five years between the two surveys. The survey will not represent any newer households that were established after the IAF was conducted.

3. Summary of Sample Design for the 2002/03 IAF

The sample for the 2002/03 IAF was based on the master sampling frame (*amostra mãe*) developed from the 1997 Mozambique Census of Population and Housing data and cartographic materials. A stratified multi-stage sample design was used. The sampling frame was stratified by province, urban and rural. The urban stratum within each province was divided into separate substrata for the capital city and other urban, and the major cities were further subdivided into socioeconomic strata defined from the 1997 Census data. The primary sampling units (PSUs) within each stratum and substratum were ordered by geographic codes to provide further implicit stratification based on the systematic selection of PSUs with probability proportional to size (PPS). Within each stratum a subsample of the census enumeration areas (EAs) in the master sample were selected for the IAF. Given that the IAF results were required at the provincial level, the sample was allocated equally to all provinces. The 78 sample EAs in each province were allocated proportionately between the urban and rural strata. Table 1 presents the final distribution of the sample EAs by province, urban and rural stratum, for the 2002/03 IAF.

The data collection for IAF was conducted over the 12-month period from July 2002 to June 2003. Given the importance of representing seasonality in household income, expenditures and other characteristics geographically throughout the year, INE attempted to assign the urban and rural sample EAs within a province to each month in a representative manner. It was also necessary to take into account logistical and operational issues in the assignment of sample EAs by month within each province, but it is expected that a representative subsample of EAs was assigned each quarter.

A listing of households was conducted in each sample EA for the 2002/03 IAF. At the last sampling stage 12 households were selected using random systematic sampling with equal probability within each sample urban EA, and 9 households within each sample rural EA. A sample of 4 potential replacement households was also selected within each sample EA in case it was necessary to replace a non-interview.

Table 1. Allocation of Sample EAs and Households by Province, Urban and Rural Stratum, for 2002/03 IAF

Province	Total		Urban		Rural	
	Sample EAs	Sample Households	Sample EAs	Sample Households	Sample EAs	Sample Households
Niassa	78	816	38	456	40	360
Cabo Delgado	78	738	12	144	66	594
Nampula	78	756	18	216	60	540
Zambézia	78	735	11	132	67	603
Tete	78	756	18	216	60	540
Manica	78	816	38	456	40	360
Sofala	78	795	31	372	47	423
Inhambane	78	756	18	216	60	540
Gaza	78	786	28	336	50	450
Maputo Province	78	837	45	540	33	297
Maputo City	78	936	78	936	-	-
Mozambique	858	8,727	335	4,020	523	4,707

4. Sampling Procedures for the PSIA Survey

The PSIA Survey is based on a subsample of the 2002/03 IAF sample, so the stratification and first two stages of selection are the same as those described for the IAF. A subsample of EAs was selected within each stratum for the PSIA Survey, and all the IAF sample households in these EAs will be screened to determine their eligibility for the survey. Two sampling alternatives were considered for selecting the subsample of EAs for the PSIA:

(1) The sample EAs assigned to a particular set of months for the 2002/03 IAF can be selected for the PSIA, and the interviews for all the IAF sample households in each of these sample EAs could be conducted during the same month in 2008. The main advantage of this sampling alternative is that it reduces the effect of seasonality in the longitudinal analysis when comparing the 2002/03 IAF data to the 2008 PSIA Survey data for the same sample households. Given the nature of the 2002/03 IAF sample design, each province would have a similar sample size for the PSIA regardless of the size of the province, although the survey results will only be produced at the national, urban and rural areas. This will result in more variable weights for the survey data between provinces, thus decreasing the statistical efficiency of the sample design.

(2) The second sampling alternative would be to allocate the subsample of EAs from the 2002/03 IAF to each province proportionally to its size (population or number of households) in order to reduce the variability in the weights. This would also make it possible to ensure a representative geographic

distribution of the sample EAs within each stratum. Although this sample would be statistically more efficient, the sample households in EAs selected for the PSIA under this approach may be interviewed in a month different from that for the 2002/03 IAF interview.

The first sampling approach was used for selecting the subsample of EAs for the PSIA. The sample EAs with households interviewed from March to May 2003 were selected for this purpose. Originally it was planned to interview all the IAF sample households in these EAs during the same month in which they had been interviewed for the 2002/03 IAF. However, because of delays in the survey planning process, the data collection for the PSIA Survey was postponed to later in 2008.

There were a total of 223 sample EAs with at least one household interviewed for the IAF between March and May 2003. The names of the heads of household for most of the IAF sample households within these sample EAs were obtained from the 2002/03 IAF listing information. However, the names could not be found for 23 sample households within these 223 sample EAs, including all the households in two sample EAs; these households are excluded from the PSIA Survey since it would not be possible to locate them again. The effective sample size for the PSIA survey, after excluding the sample households without names, is 2,234 households in 221 sample EAs. Table 2 shows the distribution of sample EAs and sample households by province and stratum for the PSIA survey.

Table 2. Distribution of Subsample of EAs and IAF Sample Households for the PSIA Survey by Province, Urban and Rural Stratum

Province	Total		Urban		Rural	
	Sample EAs	Sample Households	Sample EAs	Sample Households	Sample EAs	Sample Households
Niassa	21	222	11	132	10	90
Cabo Delgado	19	176	2	24	17	152
Nampula	21	200	4	48	17	152
Zambézia	20	188	3	36	17	152
Tete	21	207	6	72	15	135
Manica	21	222	11	132	10	90
Sofala	20	209	10	119	10	90
Inhambane	21	191	1	12	20	179
Gaza	18	183	7	84	11	99
Maputo Province	18	192	10	120	8	72
Maputo City	21	244	21	244	-	-
Mozambique	221	2,234	86	1,023	135	1,211

An attempt was made to contact all of the IAF sample households in the PSIA subsample of EAs to determine their eligibility for the survey in terms of having school-age children or older children attending school. The final sample of households for the PSIA survey was based on this screening.

During the implementation of the PSIA in the field, the IAF sample households in some sample EAs could not be found or had access problems, so it was necessary to replace 10 of the original sample EAs from the IAF March-May 2003 enumeration period. Each replacement EA was selected from the additional IAF sample EAs for the same province, urban/rural stratum as the original sample EA. Table 3 identifies the 10 original and replacement sample EAs. It was necessary to use the information from the sampling frame for the replacement EAs for calculating the weights, as described in the next section.

Table 3. Replacement of Original Sample EAs for PSIA

Province	Stratum	District	Original IAF EA No.	Replacement IAF EA No.
01 - Niassa	1 - Urban	02 – Cuamba	30	23
07 - Sofala	2 - Rural	08 – Gorongoza	501	502
07 - Sofala	1 - Urban	01 – Beira	480	489
07 - Sofala	1 - Urban	01 – Beira	481	491
10 - Maputo Province	1 - Urban	01 - C. de Matola	738	714
10 - Maputo Province	2 - Rural	06 – Matutuine	771	768
11 - Cidade de Maputo	1 - Urban	03 - M3	783	835
11 - Cidade de Maputo	1 - Urban	01 - M1	844	849
11 - Cidade de Maputo	1 - Urban	01 - M1	846	851
11 - Cidade de Maputo	1 - Urban	01 - M1	843	786

In order to determine the level of precision that can be expected for the survey estimates of key indicators from this sample size and allocation, a simulation study was carried out using the 2002/03 IAF data for a subsample of 3 months (October to December 2002). This study is described in the last section of this report. It was found that the level of precision was satisfactory for most indicators by urban and rural stratum. The number of observations for secondary students was low as expected, especially for the rural stratum, so the analysis of this domain will be more limited.

5. Weighting Procedures for PSIA Survey

In order for the sample estimates from the PSIA survey to be representative of the population, it will be necessary to multiply the data by a sampling weight, or expansion factor. The basic weight for each sample household would be equal to the inverse of its probability of selection (calculated by multiplying the probabilities at each sampling stage). The PSIA Survey is based on a subsample of the EAs selected for the 2002/03 IAF, and all the in-scope IAF sample households within these EAs will be included in the PSIA survey. Therefore the final weight for the IAF sample households will be one component of the PSIA weights. The basic IAF weights were calculated using the following formula:

$$W_{IAFhi} = \frac{M_h \times M'_{hi}}{n_h \times M_{hi} \times m_{hi}},$$

where:

W_{IAFhi} = basic weight of the IAF sample households in the i-th sample EA in stratum h

M_h = total number of households in the frame for stratum h (based on the 1997 Mozambique Census data)

M'_{hi} = total number of households listed in the i-th sample EA in stratum h

n_h = number of sample EAs selected in stratum h for IAF

M_{hi} = total number of households in the frame for the i-th sample EA in stratum h (based on the 1997 Mozambique Census data)

m_{hi} = number of sample households selected in the i-th sample EA in stratum h for IAF (12 households in urban EAs and 9 households in rural EAs)

This basic weight for the 2002/03 IAF sample households was adjusted for non-interviews and calibrated to the projected population by province for mid-2003. The final weight for each sample EA ($W"_{IAFhi}$) was then attached to the 2002/03 IAF data file for the analysis. Given that the longitudinal sample for the PSIA represents the households in Mozambique existing at the time of the 2002/03 survey, the final IAF weight will be the first component of the weight for the PSIA sample households. The second component of the PSIA weight will be based on the inverse of the subsampling rate within the corresponding stratum.

Although the original plan for the 2002/03 IAF was to have a geographically representative sample assigned to each quarter and month, the actual assignments were somewhat arbitrary, influenced by logistical considerations. It is still assumed that the geographical distribution of the sample over the months is sufficiently representative for the purposes of the PSIA analysis. The PSIA weighting procedures are based on the assumption that EAs interviewed for IAF from March to May 2003 and selected for the PSIA Survey within each stratum were a random subsample of all the IAF sample EAs for that stratum. Therefore the basic PSIA weight for the households in each sample EA can be calculated as follows:

$$W_{PSIAhi} = W''_{IAFhi} \times \frac{n_{IAFh}}{n_{PSIAh}},$$

where:

W_{PSIAhi} = weight of the PSIA sample households in the i-th sample EA in stratum h

n_{IAFh} = number of sample EAs selected in stratum h for the 2002/03 IAF

n_{PSIAh} = number of EAs selected in the subsample for PSIA in stratum h

Since the master sampling frame used for selecting the sample EAs for the 2002/03 IAF includes separate strata for the capital city and other urban areas in most provinces, as well as socioeconomic substrata for major cities, some small substrata have 0 or 1 EAs selected in the PSIA subsample. For purposes of calculating the weights it was necessary to collapse such substrata within the urban stratum of the same province. In the case of Inhambane Province, only one urban EA appears in the PSIA sample, so it is necessary to collapse the urban and rural strata for this province for the purposes of calculating the weights and variances.

It was also necessary to adjust the weights to take into account the 2002/03 IAF sample households that are excluded because the names of the heads of household could not be found. There were 23 IAF sample households without names, including two entire sample rural EAs that did not have any names available for the 18 sample households (that is, 9 households each). As a result, the two sample EAs without names were dropped from the sample. Since they are excluded from the denominator of the second component of the PSIA weight (n_{PSIAh}), the weights for the corresponding strata are automatically adjusted to take into account these two missing EAs. The remaining five households without names are in different sample EAs, so the weights for these EAs have to be adjusted accordingly. This weight adjustment can be expressed as follows:

$$W'_{PSIAhi} = W_{PSIAhi} \times \frac{m_{IAFhi}}{m'_{IAFhi}},$$

where:

m_{IAFhi} = number of sample households with completed interviews for the 2002/03 IAF in the i-th sample EA in stratum h

m'_{IAFhi} = number of interviewed sample households with the name of the head of household available in the i-th sample EA in stratum h (that is, excluding any sample households without names)

This adjustment factor is equal to 1 for all sample EAs except for the five EAs which each have one sample household without a name.

The final weights for the 2002/03 IAF data had been adjusted using population projections (based on demographic analysis) by province for the mid-point of the survey. In the case of the weights for the PSIA Survey, it is also recommended to calibrate the weights to the population projections. First it will be necessary to estimate the weighted 2003 total population from the IAF sample households in the PSIA sample EAs, in order to obtain the denominator of the ratio adjustment factors. This was done by applying the

PSIA weights to the IAF data on the number of persons for the all the 2,234 sample households selected for the PSIA (regardless of eligibility). Since only the sample households meeting the eligibility criteria for the PSIA Survey will be interviewed, the weighted total population from the final PSIA data will represent a subset of the total population. The weight adjustment factor based on the population projections by stratum can be expressed as follows:

$$A_h = \frac{\hat{P}_{03h}}{\sum_{i \in h} \sum_{j} W'_{hij} \times p_{hij}} ,$$

where:

A_h = adjustment factor for the weights in stratum h

$\hat{P}_{03h}$ = projected population of stratum h for mid-2003, based on demographic analysis

p_{hij} = number of persons in the 2002/03 IAF data for the j-th sample household in the i-th PSIA sample EA in stratum h

INE has produced population projections for the mid-point of each year by province, based on demographic analysis. The mid-2003 population projections can be used for this calibration of the weights, given that the PSIA sample corresponds to IAF sample households interviewed between March and May 2003. Since these population projections were not available by urban and rural strata within province, the weights for the 2002/03 IAF had been adjusted at the provincial level. At first this same type of provincial-level adjustment procedure was carried out for the PSIA weights. However, when the percent urban population within each province from the weighted PSIA data was compared to the corresponding results for the IAF, some provinces had considerable differences in the percent urban population; this was due to the uneven distribution of the PSIA subsample by stratum in some provinces. The difference in the percent urban population was considerable for Inhambane Province given that it only has one urban EA in the PSIA sample. In order to correct for such distortions in the distribution of the population by stratum, the 2002/03 IAF data were used to estimate the percent population by urban and rural strata within each province. These percentages were then applied to the mid-2003 population projections by province in order to estimate the population projections by urban and rural strata.

Table 4 shows the weighted population distribution from the 2002/03 IAF data with the percent population by urban and rural stratum within each province. Table 5 shows the mid-2003 projected population allocated by urban and rural stratum based on the percentages in Table 4, as well as the weighted estimates of total population by stratum using the IAF data on the number of persons in the 2,234 households in the PSIA subsample with the preliminary PSIA weights (W'_{PSIAhi}), and the corresponding weight adjustment factor for each stratum. It can be seen in Table 5 that the population weight

adjustment factors for the urban strata vary from 0.6215 for Nampula to 6.0698 for Inhambane, and for the rural strata these factors vary from 0.8595 for Inhambane to 1.4002 for Sofala. As mentioned previously, one reason for the relatively high adjustment factor for the Inhambane Urban stratum is that there is only one EA in the PSIA sample for this stratum. As a result, it was combined with the Inhambane Rural stratum for calculating the basic design weight. Since the estimate of total population for the Inhambane Urban stratum using the preliminary PSIA weights was much smaller than the corresponding population projection, the weight adjustment factor is relatively large.

Table 4. Weighted Distribution of Total Population by Province, Urban and Rural Stratum, Based on the 2002/03 IAF Data

Province	Projected Total Population Mid-2003	Weighted Population Estimate, 2002/03 IAF				
		Total Population	Urban		Rural	
			Population	Percent	Population	Percent
Niassa	941,195	941,195	180,611	19.2%	760,584	80.8%
Cabo Delgado	1,556,788	1,556,801	350,706	22.5%	1,206,095	77.5%
Nampula	3,485,420	3,485,410	1,405,043	40.3%	2,080,367	59.7%
Zambézia	3,559,923	3,559,918	386,098	10.8%	3,173,820	89.2%
Tete	1,424,263	1,424,261	215,257	15.1%	1,209,004	84.9%
Manica	1,243,638	1,243,636	448,789	36.1%	794,847	63.9%
Sofala	1,548,748	1,548,758	618,391	39.9%	930,367	60.1%
Inhambane	1,363,596	1,363,593	301,163	22.1%	1,062,430	77.9%
Gaza	1,299,521	1,299,520	333,578	25.7%	965,942	74.3%
Maputo Province	1,039,321	1,039,326	639,900	61.6%	399,426	38.4%
Maputo City	1,058,833	1,058,842	1,058,842	100.0%	-	0.0%
Mozambique	18,521,246	18,521,260	5,938,378	32.1%	12,582,882	67.9%

Table 5. Projected Mid-2003 Population by Province, Urban and Rural Stratum, with Corresponding Weighted Estimates of Total Population from the 2002/03 IAF Data for PSIA Sample Households Using Basic PSIA Weights, and Weight Adjustment Factors by Stratum

Province	Urban			Rural		
	Projected Population Mid-2003	PSIA Weighted Population	Weight Adjustment Factor	Projected Population Mid-2003	PSIA Weighted Population	Weight Adjustment Factor
Niassa	180,611	149,334	1.2094	760,584	590,337	1.2884
Cabo Delgado	350,703	190,317	1.8427	1,206,085	970,109	1.2432
Nampula	1,405,047	2,260,773	0.6215	2,080,373	1,855,675	1.1211
Zambézia	386,099	424,653	0.9092	3,173,824	2,488,175	1.2756
Tete	215,257	205,358	1.0482	1,209,006	1,381,192	0.8753
Manica	448,790	532,786	0.8423	794,848	627,123	1.2675
Sofala	618,387	706,599	0.8752	930,361	664,471	1.4002
Inhambane	301,164	49,617	6.0698	1,062,432	1,236,106	0.8595
Gaza	333,578	263,754	1.2647	965,943	764,353	1.2637
Maputo Province	639,897	704,443	0.9084	399,424	337,414	1.1838
Maputo City	1,058,833	1,107,430	0.9561	-	-	-
Mozambique	5,938,366	6,595,064		12,582,880	10,914,955	

The adjusted weight for the PSIA sample households in each sample EA can be expressed as follows:

$$W''_{PSIAhi} = W'_{PSIAhi} \times A_h,$$

where:

W''_{PSIAhi} = adjusted weight for the PSIA sample households in the i-th sample EA in stratum h

The PSIA Survey only represents the 2003 households that still live in the same location or district. Given the nature of a longitudinal survey, there are a considerable number of sample households that no longer exist in the sample EAs because they moved or the housing unit was destroyed, as well as other households without school-age children, considered to be out-of-scope for the PSIA survey. Therefore the weighted total number of households from the PSIA survey data will represent a subset of all the households in Mozambique existing in 2003. This results from the fact that the sampling frame will not represent any new households that came into existence after the IAF 2002/03 survey.

However, there were also some eligible 2002/03 IAF sample households that could not be interviewed for the PSIA because of refusals or no respondent was available. Therefore it was necessary to adjust the weights for the in-scope non-interviews. An accounting was made of the interview status for all the 2002/03 IAF sample households in the PSIA sample EAs. The first step was to determine whether each non-interview household still

exists, and whether it is in-scope for the PSIA survey. Table 6 shows the distribution of the sample households by interview status and reason for non-interview, based on the final PSIA data set. The "other" category for non-interview was specified and the questionnaire and recoded to produce this table.

Table 6. Distribution of PSIA Sample Households by Interview Status

Interview Status	No. Of Households
Completo	1,569
Recusou	5
Indisponível	10
Ausente	68
Difícil acesso	1
Doente mental	2
Faleceu	35
Fora da aldeia	1
Fora do distrito	92
Mudou de residencia	1
Não eligivel	36
Não identificado	350
Não localizado	64
Total	2,234

Regarding the non-interview categories, the first five categories (recusou, indisponível, ausente, difícil acesso and doente mental) were considered in-scope households, while the remaining categories were considered out-of-scope. The last two categories (not identified and not found) were discussed with the survey staff, and it was determined that the corresponding households probably no longer exist in the same area given that a comprehensive search was made to find them. Based on this classification of household eligibility, the final distribution of the PSIA eligible and interviewed sample households is presented in Table 7. In the case of split households, only the primary IAF sample household is counted in this table. The difference between the total number of sample households selected for the PSIA shown in Table 2 (2,234) and the number of eligible sample households in Table 3 represents households considered to be out-of-scope because they no longer exist in the same area, or they did not have any school-age children in the reference period.

Table 7. Distribution of PSIA Eligible Sample Households and Completed Household Interviews by Province, Urban and Rural Stratum

Province	Total		Urban		Rural	
	Eligible Sample Households	Interviewed Sample Households	Eligible Sample Households	Interviewed Sample Households	Eligible Sample Households	Interviewed Sample Households
Niassa	169	155	91	87	78	68
Cabo Delgado	126	126	9	9	117	117
Nampula	173	154	39	34	134	120
Zambézia	158	153	26	26	132	127
Tete	170	148	57	53	113	95
Manica	153	149	90	88	63	61
Sofala	127	124	68	67	59	57
Inhambane	165	164	11	11	154	153
Gaza	155	152	68	67	87	85
Maputo	101	95	61	57	40	38
Maputo City	158	149	158	149	0	0
Mozambique	1,655	1,569	678	648	977	921

The non-interview adjustment for the PSIA household weight within each sample EA involved determining the number of IAF sample households that were eligible for the PSIA Survey, and the number of completed interviews for the EA. The final PSIA weight was then be adjusted for non-interviews as follows:

$$W'''_{PSIAhi} = W''_{PSIAhi} \times \frac{m_{PSIAhi}}{m'_{PSIAhi}},$$

where:

m_{PSIAhi} = number of IAF sample households considered eligible for the PSIA Survey in the i-th sample EA in stratum h (including eligible non-interviews)

m'_{PSIAhi} = number of sample households with completed PSIA interviews in the i-th sample EA in stratum h

The information from the sampling frame for all the components of the basic weight for each of the 221 EAs in the PSIA sample was compiled in a spreadsheet with formulas for calculating the PSIA weights. All the sample households and persons within a sample EA will have the same final weight. Following the data collection for the PSIA Survey the information needed for the final weight adjustment for non-interviews was also entered into this spreadsheet in order to calculate the final weights to be attached to the PSIA data file.

Table 8 shows the distribution of the weighted PSIA primary sample households (excluding the additional split households) based on the final weights, by province and urban/rural stratum, and the corresponding percentage of the 2003 IAF weighted total number of households. With this table it is possible to identify the provinces and strata with the highest attrition rates. At the national level the PSIA sample represents 78.9 percent of the 2003 households. As expected, the rural areas are more stable, with a coverage of 83.1 percent compared to 69.9 percent for the urban areas. The weighted estimates to be tabulated from the PSIA Survey data will be relative values such as proportions and averages, but the weights will ensure that each stratum is correctly represented in the estimates based on the relative size of the stratum.

Table 8. PSIA Weighted Estimates of Total Number of Households by Province and Urban/Rural Stratum, and Corresponding Percent of 2002/03 IAF Weighted Estimates

Province	Total		Urban		Rural	
	PSIA Weighted Total Households	Percent of 2002/03 IAF Households	PSIA Weighted Total Households	Percent of 2002/03 IAF Households	PSIA Weighted Total Households	Percent of 2002/03 IAF Households
Niassa	158,401	82.3%	23,399	63.2%	135,002	86.8%
Cabo Delgado	303,481	69.7%	37,531	37.0%	265,950	79.7%
Nampula	794,265	85.9%	360,535	82.4%	433,730	89.0%
Zambézia	648,017	84.5%	54,129	72.9%	593,888	85.8%
Tete	259,406	83.3%	34,229	81.2%	225,177	83.6%
Manica	149,976	67.0%	43,149	58.1%	106,827	71.4%
Sofala	172,745	65.4%	66,939	58.9%	105,806	70.3%
Inhambane	244,781	87.4%	53,432	91.7%	191,349	86.3%
Gaza	232,146	86.5%	51,884	80.7%	180,262	88.4%
Maputo	102,950	53.4%	58,425	53.2%	44,525	53.6%
Maputo City	110,729	66.6%	110,729	66.6%	0	-
Mozambique	3,176,897	78.9%	894,381	69.9%	2,282,516	83.1%

6. Calculation of Standard Errors and Confidence Intervals for PSIA Estimates

For the analysis and publication of the PSIA Survey results, it is important to measure the sampling errors, confidence intervals and design effects. The sampling error of an estimate is measured by the standard error, or square root of the variance of the estimate. The variance estimator should take into account the stratification and clustering in the sample design. The design effect is defined as the ratio between the variance of an estimate based on the actual complex sample design and the corresponding variance from a simple random sample of the same size. Some statistical software packages such as Stata and SPSS include a module with a standard linearized Taylor series variance estimator that takes into account the stratification and clustering in the sample design. Another software package that can be used for producing tables of standard errors and design effects using a similar variance estimator is CENVAR, a component of the

Integrated Microcomputer Processing System (IMPS), which can be downloaded for free from the www.census.gov website.

Given that the Stata software will be used for much of the survey analysis, the svydesign feature of Stata can be used for the tabulation of sampling errors and design effects based on the actual sample design. The Stata variance estimator uses the following formula for calculating the variance of a weighted total:

Variance Estimator of a Total

$$V(\hat{Y}) = \sum_{h=1}^{L} \left[\frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \right],$$

where:

n_h = number of sample EAs selected in stratum h for the PSIA Survey

$$\hat{Y}_{hi} = \sum_{j=1}^{m_{hi}} W'''_{PSIAhi}\, y_{hij}$$

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}$$

The survey estimate of a ratio is defined as follows:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}},$$

where $\hat{Y}$ and $\hat{X}$ are estimates of totals for variables y and x, respectively, calculated as specified previously.

The variance estimator of a ratio used by Stata can be expressed as follows:

Variance Estimator of a Ratio

$$V(\hat{R}) = \frac{1}{\hat{X}^2} \left[V(\hat{Y}) + \hat{R}^2 V(\hat{X}) - 2\, \hat{R}\, COV(\hat{X}, \hat{Y}) \right],$$

where:

$$COV(\hat{X},\hat{Y}) = \sum_{h=1}^{L}\left[\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}\left(\hat{X}_{hi} - \frac{\hat{X}_h}{n_h}\right)\left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h}\right)\right]$$

$V(\hat{Y})$ and $V(\hat{X})$ are calculated according to the formula for the variance of a total.

In order to use the Stata software to calculate the standard errors and other measures of precision for the PSIA Survey, it will be necessary to specify unique stratum and PSU (EA) codes in the svydesign command. The calculation of variances requires at least two sample PSUs per stratum, so it will be necessary to combine or "collapse" any stratum with one sample PSU. As mentioned previously, there are several urban strata with only one EA in the PSIA sample, corresponding to individual capital cities or socioeconomic substrata. These substrata can be collapsed using the same criteria as those used to combine the substrata for calculating the weights, described previously. A practical way to collapse the strata is by recoding the original stratum number for one of the strata being collapsed. In the case of Inhambane, there is only one urban EA in the PSIA sample, so the same code can be used for the urban and rural strata to combine them into one collapsed stratum for the entire province. The enumeration area code (ea2) in the data file can be used as the PSU variable in the Stata specifications.

7. Simulation Study of Measures of Precision for PSIA Estimates Using IAF Data

A simulation study to estimate the level of precision for various indicators was carried out using the 2002/03 IAF data for a subsample of 3 months (October to December 2002). This subset of IAF data had been used based on the original plan to conduct the PSIA study beginning in October 2007, but the level of precision should be similar to the results using the IAF data for March to May 2003. A Stata file with IAF 2002/03 data was obtained from the World Bank, and the data for the subsample of EAs enumerated between October and December 2002 was extracted for tabulating the standard errors, confidence intervals and design effects for estimates of the net and gross primary and secondary enrolment rates, at the national level and by urban and rural strata.

The results from this simulation study are presented in Annex 1. It can be seen in these tables that the level of precision varies by indicator and the level of disaggregation. These results illustrate the confidence intervals that can be expected from a sample size similar to that of the PSIA Survey, although it should be noted that the attrition of sample households that no longer exist and other non-interviews will decrease the effective sample size. The relatively low number of observations for secondary students will limit the analysis of the corresponding indicators to the national level.

Annex 1

Tables of school enrollment rates at national level, urban and rural strata, based on IAF 2002/03 sample for October, November and December 2002, with corresponding standard errors, confidence intervals and design effects

1. IAF estimates of net enrollment for all school age children (primary and secondary - EP1, EP2, ES1 and ES2)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.621	.019	.584	.658	.030	5.503	11,555
Urban	.777	.025	.728	.826	.032	4.218	5,731
Rural	.546	.023	.500	.592	.043	5.392	5,824
Male	.651	.023	.606	.696	.035	4.325	5,515
Female	.588	.021	.548	.629	.035	3.091	6,040

2. IAF estimates of net enrollment for primary school (EP1 and EP2)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.632	.019	.594	.670	.031	3.884	11,555
Urban	.810	.026	.760	.861	.032	3.052	5,731
Rural	.556	.023	.510	.602	.042	3.713	5,824
Male	.643	.022	.599	.687	.035	2.686	5,515
Female	.620	.024	.573	.668	.039	2.859	6,040

3. IAF estimates of net enrollment for secondary school (ES1 and ES2)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.054	.012	.030	.077	.222	3.548	11,555
Urban	.133	.026	.082	.184	.194	2.684	5,731
Rural	.007	.003	.001	.014	.455	1.237	5,824
Male	.060	.013	.034	.086	.219	2.029	5,515
Female	.047	.013	.021	.073	.277	2.277	6,040

4. IAF estimates of net enrollment for first level primary (EP1)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.575	.023	.529	.620	.040	3.873	11,555
Urban	.792	.032	.729	.854	.040	3.146	5,731
Rural	.488	.026	.437	.539	.053	3.450	5,824

5. IAF estimates of net enrollment for second level primary (EP2)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.053	.012	.029	.077	.225	1.745	11,555
Urban	.121	.033	.056	.187	.274	2.157	5,731
Rural	.018	.006	.006	.030	.341	.876	5,824

6. IAF estimates of net enrollment for first level secondary (ES1)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.032	.017	-.002	.066	.545	7.406	11,555
Urban	.092	.047	.000	.184	.510	6.658	5,731
Rural	.002	.002	-.002	.005	.998	.873	5,824

7. IAF estimates of net enrollment for second level secondary (ES2)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.003	.002	.000	.006	.513	.422	11,555
Urban	.008	.004	.000	.015	.518	.434	5,731
Rural	.000	.000	.000	.000	.	.	5,824

8. IAF estimates of gross enrollment for first level primary (EP1)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	1.118	.045	1.029	1.207	.040	3.362	11,555
Urban	1.346	.050	1.249	1.444	.037	1.222	5,731
Rural	1.026	.056	.916	1.137	.054	3.695	5,824

9. IAF estimates of gross enrollment for second level primary (EP2)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.584	.049	.488	.680	.083	1.680	11,555
Urban	1.207	.073	1.064	1.351	.060	.463	5,731
Rural	.266	.049	.171	.362	.182	2.927	5,824

10. IAF estimates of gross enrollment for first level secondary (ES1)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.249	.039	.172	.326	.156	3.919	11,555
Urban	.678	.087	.506	.851	.129	1.908	5,731
Rural	.036	.013	.010	.062	.365	2.364	5,824

11. IAF estimates of gross enrollment for second level secondary (ES2)

Domain	Ratio Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Design Effect	Unweighted Count
			Lower	Upper			
Mozambique	.090	.021	.049	.131	.230	2.195	11,555
Urban	.183	.040	.104	.262	.219	1.593	5,731
Rural	.021	.021	-.019	.061	.976	5.653	5,824