Chapter 9: SURVEY WEIGHTING AND VARIANCE ESTIMATION

INTRODUCTION

This chapter describes the methods applied to compute sampling weights and estimate variances using replicate weights. The purpose of calculating sampling weights for PISA-D Strand C is to permit inferences from youth included in the sample to the population from which they were drawn and to have the tabulations reflect estimates of the population totals. Sampling weights can be considered as estimated measures of the number of units in the target population that a completed case represents. Weighting incorporates several features of the survey, including the probabilities of selection of units in the sample and adjustments for nonresponse, and any known differences between the selected sample and the total target population. Differences between the sample and the population may arise because of sampling variability, differential response rates or coverage rates among subgroups of the population, and other types of non-sampling and response errors, such as misclassification errors.

In PISA-D Strand C, survey weighting was performed to accomplish the following objectives to the extent possible:

- to permit unbiased estimates by compensating for possible disproportionate sampling of various subgroups in the sample
- to minimise biases arising from differences between respondents and non-respondents
- to compensate for non-coverage in the sample due to inadequacies in the sampling frame or other reasons for non-coverage
- to combine the representative sample with the limited representative sample
- to bring data up to the dimensions of the population totals
- to reduce sampling errors by using auxiliary data on population characteristics that are known with a high degree of accuracy
- to facilitate the estimation of variances using the replication approach.

This chapter is organised as follows. First, we provide an overview of the weighting process for each country, along with a listing of cautions and limitations for each country sample. Then we provide a description of the weighting process for the representative sample, including a discussion of the weighting steps, treatment of different disposition codes and calculation of weighting adjustment factors. Technical details for countries with a limited representative sample are provided in the next section. This is followed with the description of the compositing approach used for producing weights for countries for which the representative and limited representative samples overlapped (in the statistical representation of the selected areas). We then describe the assignment of variance units, creation of replicate weights and variance estimation procedures. In the last section, we present a discussion of the quality control process.

SURVEY WEIGHTING PROCESS OVERVIEW ACROSS COUNTRIES

In general, weighting involves adjusting for variable probabilities of selection of each of the complete cases, and deriving adjustment factors with a focus on reducing potential bias due to nonresponse, deficiencies in the sampling frame and other complications that may arise during the sample selection process. This section provides a description of the standard weighting steps employed in the PISA-D Strand C country samples. The goal was to apply the same general weighting approach to all country samples, to the extent possible, to arrive at comparable estimates of proficiency and sampling error across countries. The steps are organised as applied to two main groups: representative and limited representative samples. Table 9.1 summarises the weighting process steps used for each country sample. Each of the sample types are described in the following sections.

Table 9.1 Summary of weighting steps

Sample type	Step	Country				
		Guatemala	Honduras	Panama	Paraguay	Senegal
Representative	Base weights	Nation	Rural	Rural	Nation	Nation
	Screener eligibility and nonresponse adjustments	X	X	X	X	X
	Youth Interview eligibility and nonresponse adjustments	X	X	X	X	X
	Calibration	X	X	X	X	X
Limited representative	Base weights	Nation	Urban	Urban	In-school grade 6 or below	NA
	Calibration	X	NA	NA	X	NA
	Compositing, calibration	X	NA	NA	X	NA

A final weight is required for all sampled persons with a completed Youth Interview (YI), as well as YI literacy-related non-respondents (LRNRs) from the representative sample. The YI LRNRs from the representative sample receive a final weight despite the lack of YI or assessment data because they were considered part of the PISA-D Strand C target population and cannot be represented by survey respondents. The following steps were included in the development of the weights for the representative sample:

- assignment of a dwelling unit (DU) base weight to each sampled household to compensate for differential probabilities of selection
- DU-level eligibility and nonresponse adjustments to reduce potential biases arising from differences between respondents and non-respondents
- assignment of a youth base weight to each sampled youth to compensate for differential probabilities of selection
- youth-level eligibility adjustment and nonresponse adjustments
- trimming to reduce the impact of large weights, if necessary
- calibration of the person weights to independent control totals to compensate for

non-coverage in the sample due to deficiencies in the sampling frame.

For the limited representative sample, base weights equal to one were initially assigned. Then a calibration to control totals occurred to help in the sample's representation. A hybrid approach was used for Guatemala and Paraguay to combine the representative and limited representative samples that overlapped in the statistical representation of the selected areas. The succeeding sections describe each of the weighting steps in detail. First, some special cautionary remarks about survey quality are given below.

Cautionary remarks

One indication of survey quality is through comparison of the sum of weights (after adjusting for nonresponse) to target population totals provided by each country. For PISA-D Strand C samples, results of this comparison show large gaps between the estimated survey totals and the target population totals provided by the countries. In general, the reason for the gaps is due to non-sampling and sampling error, but could also be due to unreliable population totals. Typically, in surveys the gap is not as large as observed in PISA-D Strand C. These results indicate the great challenge in conducting the PISA-D Strand C survey, and the challenges countries face in obtaining reliable target population totals. The following provides a summary of special situations and cautionary notes for each country.

Guatemala

The final sample is composed of a statistical combination of their representative and limited representative samples through a hybrid composite weighting approach. The sample covers the entire country, and thus has adequate representation of the nation. A portion of the sample comes from applying non-probability methods. However, the weighted proportion of the non-probability sample is small (9%). For the probability sample, the sum of weights prior to calibration is only about half of the control totals provided by the country. A better stratification might be needed for the major design strata in order to find high concentration areas of the PISA-D Strand C target population. This population represents only 3.64% of the total population (total PISA-D Strand C population is 603 959 and 2018 midyear total population is 16 581 000 based on the U.S. Census Bureau International Data Base.)

Honduras

A representative sample in rural areas achieved over 1 000 completes, and a limited representative sample in urban areas achieved 120 completes, of which 15 came from schools and 105 came from a small number of handpicked locations in one or two cities. For the representative sample, the sum of weights estimates the number of youth in the target population in rural areas only. For the limited representative sample, the sum of weights will equal the sample size in the urban areas (120) because they do not represent any population but themselves. In addition, for the probability sample in the rural areas, the sum of weights prior to calibration is only about half of the control totals provided by the country. The sum of the household weights were similar to the expected counts. The hit rates were lower than expected,

but not enough to explain the difference. It could be partly due to uncertainty in the provided control totals as the data was not directly available. The totals were arrived at through two sources from different years.

Panama

The sample of over 1 800 completes from rural and indigenous areas was selected based on probability sampling. However, the level of representativeness of this sample is unknown. The limited representative sample in urban areas includes 72 completes that came from schools (youth that were in-school grade 6 or below). For the limited representative sample, the sum of weights equals the sample size in the urban areas (72) because they do not represent any population but themselves. For Panama's representative probability sample, the sum of DU weights was 57% of the number of DUs in the rural/indigenous population as reported by the country during the probability sample unit (PSU) frame/selection. There are two main components of the loss, i) the number of DUs listed (from maps) is smaller than expected; and ii) the number of DUs selected is smaller than expected in the PSUs in the final sample. Only 312 of the selected 513 PSUs were included in the data collection (this loss was accounted for during the weighting process). In summary, the outdated maps and the fact that not all PSUs were worked caused a large underestimation of the number of DUs in the rural/indigenous areas. Ironically, the sum of the person weights after nonresponse (prior to weight calibration), overestimated the total eligible population by a factor of five. This is due to the hit rate (number of completes over the number of dwelling units sampled) in rural areas being 50% versus 2% expected, and in indigenous areas was 28% versus 7% expected. For the samples in rural and indigenous areas, the sum of weights will provide an estimate of the number of youth in the target population, however, there should be much caution expressed about its representativeness.

Paraguay

Some caution is given due to some sample coming from non-probability selection methods. A representative sample achieved 814 completes and a limited representative sample achieved 188 completes. The weighted proportion of the limited representative sample is 17%. The sum of weights from the two samples together estimates the number of youth in the target population in the country excluding two departments: Boqueron and Alto Paraguay, which account for only about 2% of the target population. For the probability sample, the sum of person weights after nonresponse (prior to weight calibration) is about 80% and 50% of the control totals provided by the country for 14- to 16-year-olds in-school grade 6 or below and 14- to 16-year-olds out-of-school, respectively. This may be due to finding a lower proportion of eligible youth in the sampled PSUs than expected, based on control totals.

Senegal

The sum of person weights after nonresponse (prior to weight calibration) for 14- to16-year-olds in the PISA-D Strand C target population is lower than the control totals provided by the country by a factor of 0.77. This is most likely due to the short data collection period (26 days), the use of outdated Census 2013 listings and the possibility that some language barrier cases may not have

been properly coded. The national statistical agency only had estimates available for the out-of-school portion of the PISA-D Strand C population, so counts for the total PISA-D Strand C population were derived and may not be accurate, especially for the breakdown by region.

SURVEY WEIGHTING PROCESS FOR REPRESENTATIVE SAMPLES

For the nonresponse adjustment, variables needed to be available for all eligible units and be related to proficiency and response propensity. The pool of potential nonresponse adjustment variables came from the sampling frame, the screener and interviewer observations.

For the calibration adjustment, all variables were required to have reliable control totals and be available for all YI respondents and LRNR cases. The quality of the data from the external sources had to exceed the quality of data from PISA-D Strand C (e.g. the mean square errors of the external estimates needed to be smaller than those of the uncalibrated estimates from the survey). The concepts, definitions and coverage of the data (counts) from the external sources needed to be the same as those employed by PISA-D Strand C. Additionally, the year of the control totals needed to be as close to the data collection period as possible, ideally covering the same time period as the field period. All said, even though countries supplied their best totals available, there is still uncertainty surrounding the control totals for this challenging target population. After a thorough review, it was decided to calibrate the survey weights to the provided control totals, which are thought to have a lower mean square error than the survey estimates of the target population.

Variables used for nonresponse adjustment and calibration must have less than 5% missing data (Technical Standard 1.3). If the amount of missing data of the variables used in weighting adjustments did not exceed the 5% threshold, imputed values were generated for missing data.

Dwelling-unit-level weighting adjustments

This section outlines the weighting process at the dwelling unit (DU) level, including the creation of the DU base weights (reflecting) the DU selection probability), adjustments for PSU nonresponse, unknown eligibility status and nonresponse to the screener.

Dwelling unit base weights

The DU base weight was assigned to all sampled DUs and was computed as the reciprocal of the DU selection probability. The DU selection probability corresponded to the product of the Primary Sampling Unit (PSU) selection probability and the DU selection probability. The computation of the DU base weight is as follows:

$$W_{ij} = \frac{1}{P_{hi}P_{j|hi}},$$

where P_{hi} is the probability of selecting PSU *i* in stratum *h*, and $P_{j|hi}$ is the conditional probability of selecting DU *j* within PSU *i* of stratum *h*. For all countries except Panama, $P_{j|hi} = 1$. The DU selection probability also reflected any duplicate records in the sampling frame.

PSU nonresponse adjustment

Not all selected PSUs were included in the sample for various reasons across countries. To account for nonresponding PSUs (selected but not worked), weights of DUs in responding PSUs were adjusted by the following factor:

$$F^{PSU} = \frac{S_R^{PSU}}{S_R^{PSU} + S_{NR}^{PSU}}$$

where,

S_R^{PSU} = sum of the weighted ($\frac{1}{P_{hi}}$) measure of size across all responding PSUs in the stratum

S_{NR}^{PSU} = sum of the weighted ($\frac{1}{P_{hi}}$) measure of size across all nonresponding PSUs in the stratum.

Table 9.2 provides the average value of F^{PSU} across the selected DUs.

Table 9.2 Average value of F^{PSU} across the selected DUs

Country	Average value of F^{PSU}
Guatemala	High density stratum: 1.11 Low density stratum: 1.00
Honduras	High density stratum: 1.10 Low density stratum: 1.02
Panama	Region Metropolitan: 3.22 Central: 1.53 West: 5.33 Rural: 2.33 Indigenous: 1.58
Paraguay	High density urban: 1.11 High density rural: 1.04 Other: 1.00
Senegal	High density stratum: 1.14 Low density stratum: 1.00

Screener unknown eligibility adjustment

The first step involved an adjustment for unknown eligibility if the eligibility status of some DUs could not be determined. In this step, the weights of the DUs with unknown eligibility status (i.e. whether they were occupied) was distributed to cases known to be eligible, as follows:

$$F_1^{DU} = \frac{S_R^{DU} + S_{NR}^{DU} + S_I^{DU} + S_U^{DU}}{S_R^{DU} + S_{NR}^{DU} + S_I^{DU}}$$

where,

S_R^{DU} = sum of the weight ($F^{PSU}W_{ij}$) across all responding DUs

S_{NR}^{DU} = sum of the weight ($F^{PSU}W_{ij}$) across all nonresponding DUs

S_I^{DU} = sum of the weight ($F^{PSU}W_{ij}$) across all ineligible DUs

S_U^{DU} = sum of the weight ($F^{PSU}W_{ij}$) across all unknown eligibility status DUs.

Table 9.3 provides the average value of F_1^{DU} across the dwelling units (DUs) with known eligibility status. Table 9.4 provides the variables used for the eligibility adjustment cells.

Table 9.3 Average value of F_1^{DU} across the responding DUs with known eligibility status

Country	Average value of F_1^{DU}
Guatemala	1.03
Honduras	1.01
Panama	1.02
Paraguay	1.06
Senegal	1.05

Table 9.4 Variables used to form adjustment cells for DU unknown eligibility adjustment

Country	Weighting variables	Number of final cells
Guatemala	Major stratum (rural/urban areas), Minor stratum (socioeconomic development), Region, Interviewer observation of street lights	78
Honduras	Major stratum, Region, Interviewer observation of street lights	18
Panama	Region (Major strata)	5
Paraguay	Major stratum (high/low density of target population), Minor stratum (urbanisation), Region, Interviewer observation of street lights	44
Senegal	Major stratum (the 14 regions split into two groups of seven based on a cut off of 31% expected hit rate), Minor stratum (region), Urbanisation, Interviewer observation of street lights	38

Screener nonresponse adjustment

The next step in the weighting process was to adjust the unknown eligibility-adjusted weights to reduce potential bias because of nonresponse to the screener. For the screener nonresponse adjustment, the cases were divided into two categories: i) cases coded as screener literacy-related nonresponse (LRNR) (e.g. language barrier) or having a youth in the household coded as LRNR for the YI or assessment, and ii) all other cases. For the first group, the households

with a YI or assessment, LRNR youth will have their screener weights adjusted to account for screener-level LRNR cases. In the second group, non-LRNR respondent households will represent non-literacy related non-respondent households. Non-literacy-related non-respondents were likely to be similar to respondents with respect to proficiency scores. In contrast, households with language barriers were presumed to differ from responding households with respect to proficiency. Therefore, the weighting procedures adjusted the weights of the respondents to represent non-literacy-related non-respondents only.

The nonresponse adjustment was performed within cells, defined based on pre-selected weighting variables and found to be related to proficiency and response propensity. One variable in particular (interviewer observation of streetlights) was added based on the result of a special evaluation using Field Trial (FT) data. The evaluation examined whether the youth's proficiency level was correlated with any of the seven interviewer observation variables using a chi-square test. Since proficiency level was not available in the FT, we used the interviewer's response on how often they felt that the respondent understood the questions in the interview (referred to as UNDERSTOOD hereafter) as a proxy to the youth's proficiency level. Due to the small sample size in the FT, the evaluation was done for all countries together and for countries with a large sample size separately. Some categories of the interviewer observations and the UNDERSTOOD variable were also combined to ensure a large enough sample size for the chi-square test. The evaluation indicated that most of the seven interviewer observations were correlated with the UNDERSTOOD variable. However, the interviewer observation of streetlights was the only variable with a low missing value rate (below 5% for all countries). Therefore, it was the only interviewer observation variable used to form non-response adjustment cells.

Within each adjustment cell, the household unknown eligibility-adjusted weights of non-respondents were redistributed over a relatively large pool of cases (approximately 30 or more respondents). Additionally, the amount of variation in the nonresponse adjustment factors was kept to a minimum by limiting the maximum allowable nonresponse adjustment factor, which was a function of the achieved screener response rate. For this step, the weighting adjustment was computed for screener respondents within weighting cells as follows.

$$F_2^{DU} = \frac{S_R^{DU2} + S_{NR}^{DU2}}{S_R^{DU2}}$$

where,

S_R^{DU2}= sum of the weight ($W_{ij}F^{PSU}F_1^{DU}$) across all responding DUs

S_{NR}^{DU2}= sum of the weight ($W_{ij}F^{PSU}F_1^{DU}$) across all nonresponding DUs.

Table 9.5 provides the average value of F_2^{DU} across the responding households. Table 9.5 provides the variables used for the nonresponse adjustment cells.

Table 9.5 Average value of F_2^{DU} across the responding households

Country	Average value of F_2^{DU}
Guatemala	1.09
Honduras	1.07
Panama	1.08
Paraguay	1.75
Senegal	1.18

Table 9.6 Variables used to form nonresponse adjustment cells for the screener

Country	Weighting variables	Number of final cells
Guatemala	Literacy-related nonresponse, Major stratum (rural/urban areas), Minor stratum (socioeconomic development), Region, Interviewer observation of street lights	73
Honduras	Literacy-related nonresponse indicator, Major stratum, Region, Interviewer observation of street lights	26
Panama	Region (Major strata)	5
Paraguay	Literacy-related nonresponse indicator, Major stratum (high/low density of target population), Minor stratum (urbanisation), Region, Interviewer observation of street lights	45
Senegal	Literacy-related nonresponse, Major stratum (the 14 regions split into two groups of seven based on a cut-off of 31% expected hit rate), Minor stratum (region), Urbanisation, Interviewer observation of street lights	35

The final screener weight is computed as $FW_{ij}^{DU} = W_{ij}F^{PSU}F_1^{DU}F_2^{DU}$. Table 9.7 provides the average screener base weight, screener final weight and the design effect due to unequal weights (computed as 1 + relative variance of the weights).

Table 9.7 Average screener weights and design effect due to unequal weights

Country	Screener base weights		Screener final weights	
	Average among all selected DUs	Design effect	Average among screener respondents	Design effect
Guatemala	148.0	1.06	164.1	1.08
Honduras	50.3	1.11	53.9	1.11
Panama	35.5	5.45	39.6	5.08
Paraguay	63.1	5.10	77.4	5.29
Senegal	167.2	1.73	204.0	1.78

Youth-level weighting adjustments

This section describes the process of creating the youth-level weights for the representative sample. The steps include the following: the computation of youth base weights; the youth unknown eligibility adjustment (all non-literacy-related non-respondents are considered to have unknown eligibility status) designed to reduce potential nonresponse bias; and the calibration of weights to control totals and the general trimming procedure used to reduce the impact of extreme weights.

Youth base weights

The youth base weights for the probability samples account for both nonresponse to the household screener and differential within-household selection rates. In general, the youth base weight for youth *k,* in household *j,* in PSU *i,* were computed as the product of the screener nonresponse-adjusted weight and the reciprocal of the within-household youth selection probability. In Guatemala, Honduras, Panama and Paraguay, and for youth in Senegal classified as eligible during the screening, the within-household youth selection probability is equal to one, and therefore the youth base weight is equal to:

$$W_{ijk} = FW_{ij}^{DU}$$

In Senegal, because of evidence in the Field Trial of inconsistent classifications between the eligibility status of youth from the screener and self-reported classification from the Youth Interview, one-third of those classified as likely ineligible in the screener were randomly selected for the YI. Therefore, the youth base weights for Senegalese youth classified in the screener as likely ineligible is equal to:

$$W_{ijk} = FW_{ij}^{DU}/3$$

For referrals from the probability-based link-tracing approach, the base weight for each sampled person was computed as the reciprocal of the PSU selection probability, essentially assigning a within-PSU selection probability equal to 1 and retaining the PSU selection probability. Therefore, the youth base weight for the referred-to youth *k*, in Panama, is equal to:

$$W_{ijk} = \frac{1}{P_{hi}}.$$

Youth unknown eligibility adjustments

Adjustments for youth unknown eligibility was performed because the eligibility status of non-respondent sampled youths could not be determined due to lack of survey data. There were two adjustments, one for non-LRNR youth, and one for LRNR youth. The non-LRNR nonresponding youth were likely to be similar to respondents with respect to proficiency scores. For the second category (LRNR), types of LRNR include language problem, reading and writing difficulty and learning-mental disability. Sampled youth with this type of nonresponse were presumed to differ from respondents with respect to proficiency. Therefore, LRNRs received a different treatment than non-literacy-related non-respondents.

In the youth unknown eligibility adjustment for non-LRNR youth, within weighting cells, the youth base weights for the sampled persons with unknown eligibility status were distributed to cases with known eligibility status (determined from data collected in the Youth Interview). That is, the adjustment was computed as follows among the non-LRNR youth:

$$F_1^{YI} = \frac{S_R^{YI} + S_I^{YI} + S_U^{YI}}{S_R^{YI} + S_I^{YI}}$$

where,

S_R^{YI}= sum of the weight (W_{ijk}) across all responding eligible youth

S_I^{YI}= sum of the weight (W_{ijk}) across all ineligible youth

S_U^{YI}= sum of the weight (W_{ijk}) across all non-LRNR youth with unknown eligibility status.

The unknown eligibility adjustment was performed within cells that were defined based on pre-selected weighting variables that were hypothesised to be related to proficiency and to response propensity. Within each adjustment cell, the adjustment was done over a relatively large pool of cases (approximately 30 or more youth with known eligibility status). Additionally, the amount of variation in the adjustment factors was kept to a minimum by limiting the maximum allowable adjustment factor. The above adjustment factor was applied to non-LRNR youth with known eligibility status (in-scope respondents, and out-of-scope respondents). The weights for non-LRNR youth with unknown eligibility status were set equal to zero.

For the LRNR youth, the eligibility status is unknown. An adjustment factor was computed to reduce their weights by an observed proportion in the target population among all cases as follows:

$$F_1^{YI} = \frac{S_R^{YI}}{S_R^{YI} + S_I^{YI}}$$

The above adjustment factor was then applied to LRNR cases. For Senegal, the adjustment factor was computed within two weighting classes, which were formed as likely ineligible or eligible as determined by the screener.

Table 9.8 provides the average value of F_1^{YI} across the responding youth. Table 9.9 provides the variables used for the unknown eligibility adjustment cells for the non-LRNR youth.

Table 9.8 Average value of F_1^{YI} across the responding youth and LRNR youth

Country	Average value of F_1^{YI} for responding youth	Value of F_1^{YI} for LRNR youth
Guatemala	1.38	0.86
Honduras	1.20	0.86
Panama	1.07	0.96
Paraguay	1.21	0.86
Senegal	1.08	0.84[1]

Note: [1] Computed among eligible LRNR youth. There were no likely ineligible LRNR youth

Table 9.9 Variables used to form unknown eligibility adjustment cells for non-LRNR youth

Country	Weighting variables	Number of final cells
Guatemala	School attendance, Interviewer observation of street lights, Urbanisation, Region, Age	31
Honduras	School attendance, Interviewer observation of street lights, Urban/Rural, Region, Age	25
Panama	School attendance, Urban/Rural, Region, Age	12
Paraguay	School attendance, Interviewer observation of street lights, Urbanisation, Region, Household size, Age	14
Senegal	School attendance, Interviewer observation of street lights, Urbanisation, Region, Age	63

Adjustment to account for referrals in Panama

In Panama, probability-based link tracing was conducted, and thus the sample includes both youth selected through the probability-based households sample and those selected through referrals. As a result, the sum of weights for the probability sample and the referral sample is an overestimation of the target population. Therefore, an adjustment to the youth weights from probability-based households was conducted as follows.

Let the following be the estimated target population based on unknown eligibility adjusted youth weights from probability-based households (*prob*): $\widehat{N}^{prob} = \sum_{k \epsilon prob} W_{ijk} F_1^{YI}$.

Let the following be the contribution from the youth referrals (*ref*): $\widehat{N}^{ref} = \sum_{k \epsilon ref} W_{ijk} F_1^{YI}$.

The adjustment factor to reduce the youth weights from probability-based households becomes:

$$F_2^{YI} = \frac{\widehat{N}^{prob} - \widehat{N}^{ref}}{\widehat{N}^{prob}}$$

The value of F_2^{YI} for youth from probability-based households was equal to 0.99. For youth referrals, the factor is: $F_2^{YI} = 1$. For all other countries, $F_2^{YI} = 1$. This adjustment ensures that the sum of the adjusted weights from both sample youth from probability-based households and from youth referrals will equal $\widehat{N}^{prob}$.

Calibration

Typically, the next weighting step in survey weighting processes is to adjust the survey weights to match population control totals. This is conducted to address under-coverage bias, to reduce the mean square error (MSE) of estimates and to create consistency with statistics from other studies. For PISA-D Strand C however, there is a bit of uncertainty about the target population size and the quality of the existing external totals. The assumption is that the external population totals provided by countries are of higher quality than estimates produced by the survey itself. In that case, the adjustment is justified and will result in improved survey estimates (reduced MSE).

To help determine whether to calibrate the weights, we computed an external estimate of the PISA-D Strand C target population size by taking the difference between the total population and the Strand A estimated population size for 15 year olds, and multiplying it by 3 to expand to the age range for PISA-D Strand C (14- to 16- year olds). Then we compared that difference to the overall control total that the country provided and to the PISA-D Strand C estimate of the target population, to see which was closer. Table 9.10 provides those results. Given that what countries provided as control totals is closer to the external PISA-D Strand C estimate (difference between the total population and the Strand A estimates, multiplied by 3), it was decided to do the calibration.

Table 9.10 Comparison of external PISA-D Strand C estimate, overall control total, and PISA-D Strand C sum of youth weights

Country	External PISA-D Strand C estimate	Control total	PISA-D Strand C sum of youth weights
Guatemala	609 543	603 959	285 912
Honduras (rural areas)	192 537	151 141	78 057
Panama (rural and indigenous areas)	35 004	19 162	117 414
Paraguay	180 237	81 944	46 980
Senegal	718 836	580 996	434 669

The weights were benchmarked to control totals for different variables. Table 9.11 provides information about the control totals provided by the countries. Respondents who completed the YI and YI LRNR cases received a youth weight and were included in calibration. One iteration of calibration, trimming (if necessary) and recalibration was performed following the unknown eligibility adjustments. The calibration was conducted using a raking procedure (Deming and Stephan, 1940), which uses an iterative procedure to adjust the survey estimates to the known marginal totals of several categorical variables. For simplicity, we denote one iteration of the raking procedure as follows.

$$F_3^{YI} = \frac{S^*}{S_2^{YI}}$$

where,

S^* = control total for the cell

$S_2^{YI} = \sum_k W_{ijk} F_1^{YI} F_2^{YI}$, for all YI respondents and YI LRNR cases.

Table 9.11 Source, years and exclusions relating to control totals

Country	Source (provider)	Year	Exclusions
Guatemala	Ministry of Education National Statistics Institute	2018	None
Honduras	Ministry of Education National Statistics Institute	2013 2017	Urban areas
Panama	Ministry of Education	2018	Urban areas
Paraguay	Ministry of Education and Science Institute of Statistics	2018	Boqueron and Alto Paraguay departments
Senegal	Ministry of Education National Statistical Agency	2015	None

Trimming the outliers

Even a carefully designed sample could not fully prevent the need for reducing extreme weights. Sample designs that include the selection of dwelling units from strata with different selection rates (such as the design used in PISA-D Strand C) have increased variability in the weights. The use of nonresponse and calibration adjustments also introduces variations in sampling weights. Weight trimming reduces the dominance of large weights on the outcome statistics. The number of weights to be trimmed is usually kept to a small number because trimming introduces some bias into the survey estimates. However, the trimming adjustment in most cases reduces the sampling error component of the overall mean square error more than it increases the bias as the adjustment is applied to only a relatively small number of weights (Lee, 1995). After weight trimming, the weights are recalibrated to match control totals once again.

For PISA-D Strand C, the youth weights were trimmed as deemed necessary after the first calibration step. For Guatemala, Honduras and Senegal, a design-based procedure was used

where cells for trimming were formed from groups that were expected to be approximately self-weighting. In each cell, weights above a cutoff value were trimmed down to the designated cutoff. To define the trimming cutoff point, the coefficient of variation (CV) based on the weights after raking (the cut point was calculated separately by domain in case oversampling was used for some domains) was examined. The weights that were over 5 times the median weight were considered to be trimmed to the value equal to 5 times the median weight. For Panama and Paraguay, the large variation in weights shown in Table 9.7 was due to the implementation of a cost-reducing sample design with large variation in sampling rates across high and low concentration strata. Therefore, to reduce the impact of extreme weights, one trimming cell was used and the trimming cutoff point was defined as 7.5 times the median weight for Panama and 5 times the median weight for Paraguay. The trimming factor is computed as follows:

$$F_4^{YI} = \frac{cutoff}{W_{ijk}F_1^{YI}F_2^{YI}F_3^{YI}}, \text{ if } W_{ijk}F_1^{YI}F_2^{YI}F_3^{YI} > \text{cutoff}$$

Otherwise, the trimming factor is set equal to 1. The count and minimum trimming factor is provided in Table 9.12.

Table 9.12 Number of youth weights trimmed and minimum trimming factor

Country	Number of cases trimmed	Minimum trimming factor F_4^{YI}
Guatemala	1	0.93
Honduras	0	-.-
Panama	184	0.21
Paraguay	130	0.04
Senegal	60	0.33

After trimming, the weights were recalibrated back to the control totals. This last step results in a minor adjustment and so it is not included in the notation. Table 9.13 provides the list of variables used and the average adjustment factors for the rake-trim-rake process.

Table 9.13 Variables used in weight calibration and average raking and trimming factors

Country	Variables	Average rake-trim-rake factor $F_3^{YI}F_4^{YI}$
Guatemala	Urbanisation, School attendance by age, Region, Gender	Rake: 2.02 Trim: 1.00 Rake: 1.00 Overall: 2.02
Honduras	Gender, Region, School attendance	Rake: 1.91 Trim: 1.00 Rake: 1.00 Overall: 1.91
Panama	Region, School attendance	Rake: 0.49 Trim: 0.97 Rake: 1.03 Overall: 0.49

Country	Variables	Average rake-trim-rake factor $F_3^{YI} F_4^{YI}$
Paraguay	Gender, Urbanisation, School attendance	Rake: 3.15 Trim: 0.89 Rake: 5.53 Overall: 11.2
Senegal	Urbanisation, School attendance, Region, Gender	Rake: 1.63 Trim: 1.00 Rake: 1.01 Overall: 1.65

A summary of the weighting process adjustments for the representative sample is provided in Table 9.14. For Paraguay, large weight variation was due to probability proportionate to size selection of PSUs with highly different selection rates within strata, coupled with conducting a mini-census within areas.

Table 9.14 Summary of representative sample weighting process

Country	DU base weights	Youth base weights	Youth unknown eligibility adjusted weights	Youth raked weights	Youth rake-trim-raked weights
Guatemala					
Number of records	25 875	1 992	1 250	1 250	1 250
Sum of weights	3 830 084.33	330 325.23	285 911.73	603 959.00	603 959.00
Mean of weights	148.02	165.83	228.73	483.17	483.17
Minimum weight	83.12	86.45	98.75	62.87	62.82
Maximum weight	290.14	439.71	589.17	1 877.16	1 877.93
$1 + CV^2$	1.06	1.07	1.12	1.43	1.43
Honduras					
Number of records	18 582	1 626	1 161	1 161	1 161
Sum of weights	934 926.47	90 953.21	78 112.16	151 141.00	151 141.00
Mean of weights	50.31	55.94	67.28	130.18	130.18
Minimum weight	18.38	18.38	19.14	27.99	27.99
Maximum weight	59.14	66.75	99.11	267.34	267.34
$1 + CV^2$	1.11	1.09	1.11	1.16	1.16
Panama					
Number of records	6 081	2 249	1 983	1 983	1 983
Sum of weights	215 727.14	123 397.32	117 478.85	19 162.00	19 162.00
Mean of weights	35.48	54.87	59.24	9.66	9.66
Minimum weight	2.88	1.00	1.06	0.50	0.46
Maximum weight	500.19	519.58	515.78	165.95	41.06
$1 + CV^2$	5.45	4.18	4.04	2.81	2.24
Paraguay					
Number of records	28 709	1 165	814	814	814
Sum of weights	1 810 992.17	49 766.45	42 521.31	81 944.00	81 944.00
Mean of weights	63.08	42.72	52.24	100.67	100.67
Minimum weight	1.25	1.29	1.41	1.20	6.4

Country	DU base weights	Youth base weights	Youth unknown eligibility adjusted weights	Youth raked weights	Youth rake-trim-raked weights
Maximum weight	559.24	802.73	822.46	1 763.66	582.11
1 + CV²	5.10	8.95	8.78	8.20	2.26
Senegal					
Number of records	8 774	3 413	2 103	2 103	2 103
Sum of weights	1 467 155.77	1 026 781.38	434 669.22	580 996.00	580 996.00
Mean of weights	167.22	300.84	206.69	276.27	276.27
Minimum weight	42.80	45.40	38.10	23.81	23.31
Maximum weight	682.87	2 714.05	2 284.09	2 412.44	2 425.82
1 + CV²	1.73	2.59	2.03	1.94	1.89

SURVEY WEIGHTING PROCESS FOR THE LIMITED REPRESENTATIVE SAMPLE

The limited representative sample weighting process differed for two sets of countries. First, the limited representative samples in Honduras and Panama came from the urban areas whereas the representative sample was from other areas (indigenous and rural in Panama, and rural in Honduras). Second, in Guatemala and Paraguay, the representative and limited representative samples overlapped in the statistical representation of the selected areas. Senegal did not have a limited representative sample.

Honduras

Because the location (LOCA) and school administration (SCAD) samples came from a small number of hand-picked organisations and schools in urban areas, a youth final weight equal to 1 was assigned to all youth ($W0_{ijk} = 1$), and no further adjustments were made. Therefore, the youth in urban areas that completed the interview and assessment only represent themselves, while the youth in rural areas represent themselves as well as other youth in rural areas.

Panama

The youth in the limited representative sample, which occurred in urban areas, received a final weight equal to 1 ($W0_{ijk} = 1$). The SCAD sample came from a small number of hand-picked schools, and therefore, the youth in Panama's urban areas that completed the interview and assessment only represent themselves, while the youth in indigenous and rural areas represent themselves and other youth in similar areas (with caution as expressed in prior section). Table 9.15 provides a summary of the representative and limited representative sample weights for Honduras and Panama.

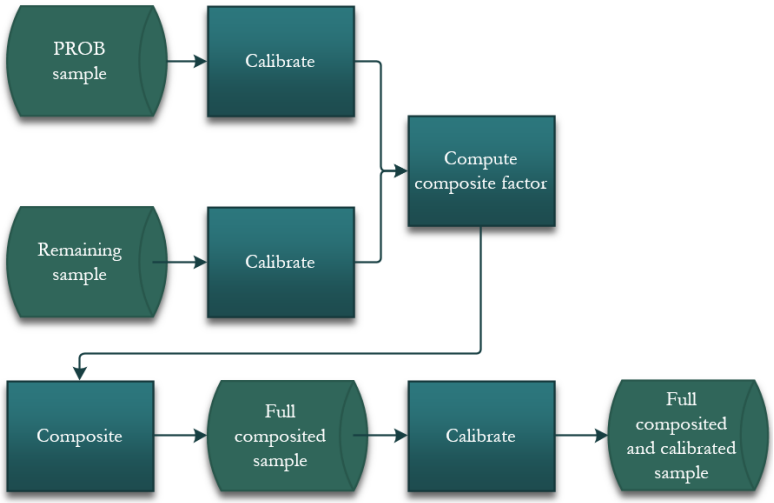
Table 9.15 Sum of sample count and final full sample weights for Honduras and Panama

Country	Limited representative sample		Representative sample				Total	
	Urban		Rural		Indigenous			
	n	Sum of weights	n	Sum of weights	n	Sum of weights	n	Sum of weights
Honduras	120	120.00	1 161	151 141.00	NA	NA	1 281	151 261.00
Panama	72	72.00	525	13 422.32	1 458	5 739.57	2 055	19 234.53

Guatemala

First, as shown in Figure 9.1, weights were calibrated to the same control totals as the representative sample. Let F_3^{YI} denote the calibration factor for the limited representative sample. Then, the calibrated weights from the whole representative sample (referred to in the chart as "PROB" for probability-based sample) were composited with weights from the whole limited representative sample. Once composited, weights were recalibrated to the original control files used for the representative sample as discussed in the prior section. Let F_5^{YI} denote the recalibration factor.

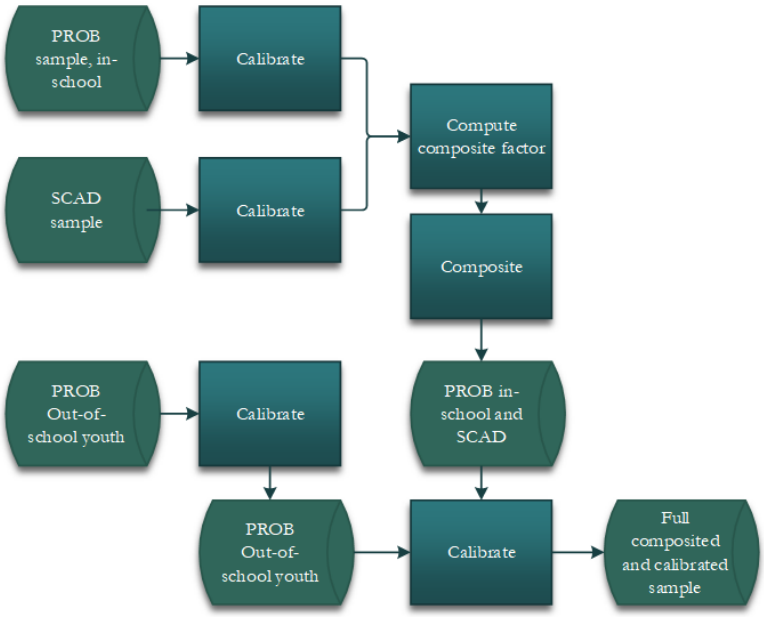
Figure 9.1 Limited representative sample weighting process for Guatemala



Paraguay

As shown in Figure 9.2, weights from the representative sample (referred to in the chart as "PROB" for probability-based sample) for in-school youth attending grade 6 or below were composited with weights from the SCAD sample. Once composited, the data were combined with the out-of-school youth sample during a final calibration step. The sample was calibrated to the original control files used for the representative sample as discussed in the prior section.

Figure 9.2 Limited representative sample weighting process for Paraguay



The goal of the composite factor computation was to weigh more heavily on the sample with trust in higher quality (higher power of representation). The quantification of that trust is in the form of the compositing factor. One approach is to use the design effect (DEFF) due to unequal weights, which is computed as $1 + cv^2$, where the *cv* is the coefficient of variation of the weights. The computation of *cv* was based on the calibrated weights for the PROB sample and of the *cv* on the calibrated weights for the limited representative sample. Along with *DEFF* due to unequal weights, the use of the Kolmogorov-Smirnov (K-S) statistic (*d*) provided a way to penalise further the non-probability sample. The statistic *d* is the greatest difference in the cumulative sample distribution of the sample compared to the cumulative control total distribution. Suppose n_1 is the sample size for the probability cases, and n_2 is the sample size for the limited representative sample cases (similar notation for design effect due to weighting, and the K-S statistic), then the compositing factor was computed as follows:

$$a = n_1/(1 + cv_1^2)/(n_1/((1 + cv_1^2) + \left(\frac{n_2}{(1 + cv_2^2)(1 + d_2)}\right)$$

The cumulative distribution for d_2 was based on the pre-calibrated weights for the limited representative sample. In this manner, the statistic d_2 was the greatest difference in the cumulative sample distribution of the limited representative sample compared to the estimated cumulative control total distribution. Essentially, the further away that the limited representative sample distribution is from the control total distribution, the less impact on the survey estimates it will have.

Table 9.16 provides a summary of the compositing and calibration steps for Guatemala and Paraguay. For Guatemala, the compositing factor was equal to 0.91, which means that much

more influence is given to the probability sample relative to the limited representative sample. For Paraguay where compositing was done for in-school 6th graders or below, the factor was equal to 0.44 (coincidently the sample sizes were equal). This gives more influence to the limited representative sample. This is an acceptable outcome because the limited representative sample resulted in a close-to-probability sample of students within schools. The schools were selected randomly, however there were limitations due to exclusions and quota sampling within schools. In the case of composite weighting, the probability sample of in-school youth attending grade 6 or below was down-weighted due to the large variation in the weights.

Table 9.16 Summary of compositing and calibration steps for Guatemala and Paraguay

Country	Sample sizes involved in compositing		Composite factor α	Average calibration factor F_3^{YI} for the limited representative sample	Average re-calibration factor F_5^{YI}, full combined sample	
	Limited representative sample	Representative sample			Sample size	Average of F_5^{YI}
Guatemala	499	1 250	0.91	1 210.34	1 759	1.00
Paraguay	188	188	0.44	89.76	1 002	1.00

FINAL WEIGHTS

Guatemala and Paraguay

The weights for representative sample cases were multiplied by α and weights for limited representative cases were multiplied by $1 - \alpha$. Therefore, the final full sample weights ($W0$) for youth *k* in the combined representative and limited representative sample for Guatemala and Paraguay were computed as:

$W0_{ijk} = W_{ijk}F_1^{YI}F_2^{YI}F_3^{YI}F_4^{YI}\alpha F_5^{YI}$, for the representative sample Guatemala and Paraguay, and

$W0_{ijk} = W_{ijk}F_3^{YI}(1-\alpha)F_5^{YI}$, for the limited representative sample in Guatemala and Paraguay.

Summary information on the last steps in combined sample weighting process for Guatemala and Paraguay is given in Table 9.17.

Table 9.17 Summary of the combined sample weights for Guatemala and Paraguay

Country	Youth composited weights	Youth re-calibrated weights
Guatemala		
Number of records	1 749	1 749
Sum of weights	603 959.00	603 959.00
Mean of weights	345.32	345.32
Minimum weight	2.13	2.13
Maximum weight	1 702.14	1 702.14
$1 + CV^2$	1.77	1.77

Country	Youth composited weights	Youth re-calibrated weights
Paraguay		
Number of records	1 002	1 002
Sum of weights	81 944.00	81 944.00
Mean of weights	81.78	81.78
Minimum weight	2.81	2.81
Maximum weight	582.11	582.11
$1 + CV^2$	2.51	2.51

Honduras, Panama and Senegal

The final full sample weight for youth *k* for the representative sample is the product of the initial youth base weight (resulting from the screener weighting process) and the youth-level adjustment factors:

$$W0_{ijk} = W_{ijk}F_1^{YI}F_2^{YI}F_3^{YI}F_4^{YI}$$

For urban areas in Honduras and Panama, the final weights were set equal to ($W0_{ijk} = 1$).

In some cases, after the adjustment, the full sample weight was less than 1. These weights were then set to equal 1. The number of cases where this occurred was as follows: Panama (4) and Paraguay (49).

Table 9.18 summarises the generalisability of estimates when using the final weights for each country. For Honduras, the use of the final weights allows generalisations in rural areas. For Panama, the use of this weight allows generalisations of results in rural and indigenous areas with cautionary remarks as provided in the introduction section. The respondents in urban areas for both Honduras and Panama only represent themselves.

Table 9.18 Generalisability of estimates

Country	Generalisability of estimates	Cautions
Guatemala	National	Large underestimate in sum of weights prior to weight calibration step. Some caution is given due to some sample coming from non-probability methods.
Honduras	Rural population, urban respondents	Large underestimate in sum of weights prior to weight calibration step occurred for rural areas. The urban sample is purposively selected and very limited.
Panama	Rural and indigenous population, urban respondents	Large overestimate in sum of weights prior to weight calibration step for rural and indigenous areas. This is likely due to unexpected extreme hit rates especially in rural areas, and fewer DUs listed and selected than expected given the data provided. The urban sample is purposively selected and very limited.
Paraguay	National	Large underestimate in sum of weights prior to weight calibration step. Some caution is given due to some sample coming from non-probability methods. Two departments were excluded: Boqueron and Alto Paraguay.

Country	Generalisability of estimates	Cautions
Senegal	National	Low-to-moderate underestimate in sum of weights prior to weight calibration step. This is likely due to a short data collection period, and use of outdated DU listings.

VARIANCE ESTIMATION

Inferences will not be valid unless the corresponding variance estimators appropriately reflect all of the complex features of the PISA-D Strand C sample design (e.g. stratification and clustering). The replication approach is used for estimating variances for the international analyses of PISA-D Strand C data. Under the replication approach, subsamples (also known as replicates) from the full sample are formed and statistics of the subsamples are used to estimate the variance of the full sample statistic. The sample replication approach, in conjunction with the multiple imputation approach used to estimate proficiency levels, captures the variation due to the complex sampling and measurement approaches, including:

- sample design
- selection
- weighting adjustments
- measurement uncertainty.

The approach used to estimate sampling variance for PISA-D Strand C was the delete-one jackknife, which is also referred to as delete-a-group jackknife, random groups approach, or JK1. Replication methods are applied to surveys by dividing the sample into specially designed replicate subsamples that mirror the design of the full sample. This is achieved by means of creating and using replicate weights.

Derivation of variance estimates for the limited representative sample required the assumption that all sample units were selected randomly.

Creation of replicate weights

The specification of variance units reflected the sample designs for each country. First, all selected DUs were sorted in sample selection order, for example, within major strata and PSU. Then the first-stage units (PSUs) were assigned sequential numbers. For example, in PSU 1, all DUs within PSU 1 were assigned variance unit 1, for PSU 2, all DUs within PSU 2 were assigned variance unit 2, and so on. For the 31st PSU, the ordering restarts with a value of 1 for the assigned variance unit, and so forth.

For Guatemala, the first-stage units for the school-based samples were the schools, and PSU was used for their location sample. For Honduras, the youth ID was used as the first-stage unit. For Panama and Paraguay, the first-stage unit for variance unit assignment was the school.

Next, 30 replicate base weights were created for each country. The DU base weights were replicated as follows. For replicate weight 1, the weights for DUs that were assigned

variance unit 1 were set to equal 0, and a factor of 30/29 was applied to the DU base weights for all other DUs. For replicate weight 2, the weights for DUs that were assigned variance unit 2 were set to equal 0, and a factor of 30/29 was applied to the DU base weights for all other DUs, and so on until all 30 replicate weights were formed. Subsequently, all weight adjustments that were conducted for the full sample were conducted on each replicate weight to capture the variation created, or reduced, by the weight adjustments.

Sampling variance estimation using replicates

Once the replicate weights are created, an estimate is then calculated for the full sample and each of the replicate subsamples. The variance of the full sample estimate is computed as the sum of squared deviations between each replicate subsample estimate and the full sample estimate. The replication formula for JK1 as applied to PISA-D Strand C is

$$Var(\hat{\theta}) = \frac{29}{30}\sum_{l}(\hat{\theta}_l - \hat{\theta}_0)^2$$

where,

$\hat{\theta}_0$ = full sample estimate

$\hat{\theta}_l$ = estimate for replicate *l.*

Variance estimation using plausible level values

When the statistic of interest involves plausible level values, the calculation above needs to be repeated separately with each of the *M* plausible level values, and the sampling variance estimate is the average of the *M* variances, as shown below.

$$SVar(\bar{\hat{\theta}}) = \sum_{m=1}^{M}\left(\frac{29}{30}\sum_{l}(\hat{\theta}_{l,m} - \hat{\theta}_{0,m})^2\right)/M$$

The estimator of the population then becomes the average of the *M* estimates calculated using each of the plausible level values, or:

$$\bar{\bar{\hat{\theta}}} = \sum_{m=1}^{M}\hat{\theta}_{0,m}\Big/M$$

Where M is the number of plausible level values and $\hat{\theta}_{0,m}$ are the statistics calculated with each of the plausible level values. The measurement variance of the estimated statistic $\bar{\bar{\hat{\theta}}}$ is computed using formulas specific to multiple imputations as follows:

$$MVar\left(\bar{\hat{\theta}}\right) = \left(1 + \frac{1}{M}\right) * \left(\sum_{m=1}^{M}\left(\hat{\theta}_{0,m} - \bar{\hat{\theta}}\right)^2\Big/M - 1\right)$$

The final variance of the statistic will be the combination of the sampling variance and the corresponding measurement variance.

$$Var(\bar{\bar{\theta}}) = SVar\left(\bar{\bar{\theta}}\right) + MVar(\bar{\bar{\theta}})$$

WEIGHTING QUALITY CONTROL CHECKS

Quality control (QC) checks were performed for both the full sample and replicate weights after each adjustment in the weighting procedure to ensure proper implementation. Performing the weighting QC checks was essential for verifying that the final weights produced for estimation are appropriate. The PISA-D Strand C schedule required the weighting QC checks to be conducted prior to the development of proficiency scores. Further checks are conducted after derivation of the proficiency scores if analyses showed any need for reverification/correction of the weights. All participating countries in PISA-D Strand C were responsible for preparing input files for weighting. The international contractor was responsible for deriving sampling weights for the Main Study for all countries.

LESSONS LEART

Based on the Field Trial and Main Survey experience of PISA-D Strand C, the international contractor (Westat) is outlining a series of lessons learnt as it relates to weighting activities:

- There was limited information collected for non-respondents. If data is available for respondents and non-respondents, and the data are related to the survey outcome and the response propensity, then the potential for bias can be reduced.
- Curbside observations are potentially useful for reducing bias due to nonresponse to the screener. The amount of missing values related to curbside observations made some variables unusable.
- Detailed instructions were provided to countries for checking the Sample Design International File (SDIF), the base file for the weighting process. In addition, extensive feedback was provided to countries on a preliminary SDIF, to prepare them for the submission of the final file. Nevertheless, many iterations of comments and subsequent corrections occurred immediately after the data collection period on the SDIF. This process took much longer than expected.
- An initial large weight variation for Panama and Paraguay reflected their sample design. To reduce cost of screening, there were very different sampling rates by density strata. The largest weights were all out-of-school youth from the low-density stratum, which was under-sampled. The set of weights were unusable for analysis purposes due to the large weight variation. Therefore, a larger than usual amount of weight trimming was conducted and the trimming classes that reflect the sample design were not used.

RECOMMENDATIONS

Based on the Field Trial and Main Survey experience of PISA-D Strand C, the international contractor (Westat) is proposing a series of weighting-related recommendations if the PISA-D Strand C population is to be incorporated in future cycles of PISA.

- The National Centre relationship with the National Statistical Institute (NSI) is crucial during the weighting process. It is critical to have access to high quality population data, usually only available through country NSIs. There was a lack of detail in control totals to conduct the weight calibration step, and uncertainty in their quality. The control totals could be multi-dimensional, including an aggregation of region, urban/rural, school enrolment status and gender. Bias due to under-coverage and non-response can potentially be reduced with good quality control totals.
- The country and the NSI can conduct the analysis of the control totals that is shown in Table 9.10 jointly. This would save time during the critical path of the weighting process that is conducted by the international contractor.
- More emphasis can be placed on gathering area-level information for small geographies (e.g. unemployment), and curbside observations, to reduce bias due to screener nonresponse.
- More emphasis could occur during training of interviewers to limit the amount of missing values related to curbside observations.
- To limit back and forth between countries and contractors for the SDIF task, a computer programme of quality control checks could be provided to the countries to run prior to loading the SDIF.

REFERENCES

Deming, W. E., & Stephan, F. F. (1940), "On a least square adjustment of a sampled frequency table when the expected marginal totals are known", *Annals of Mathematical Statistics,* 11, pp. 427-444.

Lee, H. (1995). *"*Outliers in business surveys*", In B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge & P. Kott (Eds.), Business Survey Methods* (pp. 503-526), New York, NY: John Wiley & Sons.