Chapter 8: SURVEY WEIGHTING AND THE CALCULATION OF SAMPLING VARIANCE

Survey weights are required to analyse PISA-D data, calculate appropriate estimates of sampling variance, and make valid estimates and inferences of the population. The PISA-D Consortium calculated survey weights for all students selected to participate in the survey regardless of whether they were assessed, ineligible, or excluded students, and provided variables in the data that permit users to make approximately unbiased estimates of standard errors, conduct significance tests, and create confidence intervals appropriately given the complex sample design for PISA-D implemented in each individual participating country.

SURVEY WEIGHTING

While the students included in the final PISA-D sample for a given country were chosen randomly, the selection probabilities of the students varied. Survey weights must therefore be incorporated into the analysis to ensure that each sampled student appropriately represents the correct number of students in the full PISA-D population.

There are several reasons the survey weights are not the same for all students in a given country:

- A school sample design may intentionally over- or undersample certain sectors of the school population. Oversampling would occur so certain sectors, such as a relatively small but politically important province or region, or a subpopulation using a particular language of instruction, could be effectively analysed separately for national purposes. Undersampling would occur for reasons of cost or other practical considerations such as very small or geographically remote schools.[1]
- Information about school size available at the time of sampling may not have been completely accurate. If a school was expected to be large, the selection probability was based on the assumption that only a sample of students would be selected from the school for participation in PISA-D. But if the school turned out to be small, all students would have to be included. In this scenario, the students would have a higher probability of selection in the sample than originally planned, making their inclusion probabilities higher than those of most other students in the sample. Conversely, if a school assumed to be small actually was large, the students included in the sample would have smaller selection probabilities than others.
- School non-response, where no replacement school participated, may have occurred, leading to the under-representation of students from that kind of school unless weighting adjustments were made.
- Student non-response, within participating schools, occurred to varying extents. Sampled students who were PISA-D-eligible and not excluded, but did not participate in the

assessment for reasons such as absences or refusals, would be under-represented in the data unless weighting adjustments were made.

- Trimming the survey weights to prevent undue influence of a relatively small subset of the school or student sample might have been necessary if a small group of students would otherwise have much larger weights than the remaining students in the country. Such large survey weights can lead to estimates with large sampling errors and inappropriate representations in the national estimates. Trimming survey weights introduces a small bias into estimates but greatly reduces standard errors (Kish, 1992).

The procedures used to derive the survey weights for PISA-D reflect the standards of best practice for analysing complex survey data and the procedures used by the world's major statistical agencies. The same procedures were used in PISA and other international studies of educational achievement such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Studies (PIRLS). The underlying statistical theory for the analysis of survey data as collected in PISA-D can be found in Cochran (1977), Lohr (2010), and Särndal, Swensson, and Wretman (1992).

Weights are applied to student-level data for analysis. The weight, W_{ij}, for student *j* in school *i* consists of two base weights—the school base weight and the within-school base weight—and five adjustment factors, and can be expressed as:

$$W_{ij} = (w_{1i} * f_{1i} * t_{1i})(w_{2ij} * f_{2ij} * t_{2ij})$$

where:

w_{1i} is the school base weight, is given as the reciprocal of the probability of inclusion of school *i* into the sample;

f_{1i} is an adjustment factor to compensate for non-participation by other schools that are somewhat similar in nature to school *i* (not already compensated for by the participation of replacement schools);

t_{1i} is a school base weight trimming factor, used to reduce unexpectedly large values of w_{1i};

w_{2ij} is the within-school base weight, given as the reciprocal of the probability of selection of student *j* from within the selected school *i*;

f_{2ij} is an adjustment factor to compensate for non-participation by students within the same school non-response cell and explicit stratum, and, where permitted by the sample size, within the same high/low grade and gender categories; and

t_{2ij} is a final student weight trimming factor, used to reduce the weights of students with exceptionally large values for the product of all the preceding weight components.

The school base weight

The term w_{1i} is referred to as the school base weight. For the systematic sampling with probability proportional-to-size method used in sampling schools for PISA-D, this weight is the reciprocal of the selection probability for the school, and is given as:

$$w_{1i} = \begin{cases} I_g / MOS_i & \text{if } < MOS_i < I_g \\ 1 & \text{otherwise} \end{cases}$$

The term MOS_i denotes the measure of size given to each school on the sampling frame.

The term I_g denotes the sampling interval used within the explicit sampling stratum g that contains school i and is calculated as the total of the MOS_i values for all schools in stratum g divided by the number of schools selected for that stratum.

MOS_i was set as equal to the estimated number of 15-year-old students in school i (denoted as EST_i) if it was greater than the predetermined target cluster size (*TCS*), which ranged from 40 to 42 students for the PISA-D countries. For smaller schools, the value of MOS_i is given via the following formula, where again, EST_i denotes the estimated number of 15-year-old students in the school:

$MOS_i = EST_i$, if $EST_i >= TCS$;

$= TCS$, if $TCS > EST_i >= TCS/2$;

$= TCS/2$, if $TCS/2 > EST_i > 2$;[2]

$= TCS/4$, if $EST_i = 0, 1$ or 2.

These different values of the MOS are intended to minimise the impact of small schools on the variation of the weights while recognising that the per-student cost of assessment is greater in small schools.

Thus, if school i was estimated to have one hundred 15-year-old students at the time of sample selection, $MOS_i = 100$. If the country had a single explicit stratum (g=1) and the total of the MOS_i values over all schools in the country was 150 000 students, with 150 schools selected, then the sampling interval was I_1 = 150 000/150 = 1 000 for school i (and others in the sample), giving a school base weight of $w_{1i} = 1000/100 = 10.0$. Thus, the school can be thought of as representing about 10 schools in the population. In this example, any school with 1 000 or more 15-year-old students would be included in the sample with certainty, with a base weight of $w_{1i} = 1$ as the MOS_i; this is larger than the sampling interval. In the case where one or more schools have an *MOS* value that exceeds the relevant value of *I*, these schools become certainty selections, and the value of *I* is recalculated after removing them.

The school base weight trimming factor

Once school base weights were established for each sampled school in the country, verifications were made separately within each explicit sampling stratum to determine if the school base

weights required trimming. The school trimming factor, t_{1i}, is the ratio of the trimmed to the untrimmed school base weight, and for most schools (and therefore most students in the sample) is equal to 1.0000.

The school-level trimming adjustment was applied to schools that turned out to be much larger than was assumed at the time of school sampling. Schools were flagged where the 15-year-old student enrolment exceeded 3 × *MAX*(*TCS*, MOS_i). For example, if the *TCS* was 42 students, then a school flagged for trimming had more than 126 (=3 x 42) PISA-D-eligible students, and more than three times as many students as was indicated on the school sampling frame. Because the student sample size was set at *TCS* regardless of the actual enrolment, the student sampling rate was much lower than anticipated during the school sampling. This meant that the weights for the sampled students in these schools would have been more than three times greater than anticipated when the school sample was selected. These schools had their school base weights trimmed by having MOS_i replaced by 3 × *MAX*(*TCS*, MOS_i) in the school base weight formula. This means that if the sampled students in the school would have received a weight more than three times larger than expected at the time of school sampling (because their overall selection probability was less than one-third of that expected), then the school base weight was trimmed so that such students received a weight that was exactly three times as large as the weight that was expected.

The choice of the value of 3 as the cutoff for this procedure was based on experience with balancing the need to avoid variance inflation due to weight variation that was not related to oversampling goals without introducing any substantial bias by altering many student weights to a large degree. School weights required trimming in only one country.

The within-school base weight

The term w_{2ij} is referred to as the within-school base weight. With the PISA-D procedure for sampling students, w_{2ij} did not vary across students (*j*) within a particular school *i*. That is, all of the students within the same school had the same probability of selection for participation in PISA-D. This weight is given as:

$$w_{2ij} = enr_i / sam_i$$

where enr_i is the actual enrolment of 15-year-old students in the school on the day of the assessment (and so, in general, is somewhat different from the MOS_i), and sam_i is the sample size within school *i*. It follows that if all PISA-D-eligible students from the school were selected, then $w_{2ij} = 1$ for all eligible students in the school. For all other cases, $w_{2ij} > 1$ as the selected student represents other students in the school besides themselves.

The school non-response adjustment

In order to adjust for the fact that those schools that declined to participate, and were not replaced by a replacement school, were not generally typical of the schools in the sample as a whole, school-level non-response adjustments were made. Within each country, sampled schools were formed into groups of similar schools by the international sampling and weighting

contractor. Then, within each group, the weights of the responding schools were adjusted to compensate for the missing schools and their students.

The compositions of the non-response groups varied from country to country but were based on cross-classifying the explicit and implicit stratification variables used at the time of school sample selection. It was desirable to ensure that each group had at least six participating schools because small groups could lead to unstable weight adjustments, which in turn would inflate the sampling variances. Adjustments greater than 2.0 were also flagged for review because they could have caused increased variability in the weights and led to an increase in sampling variances. It was not necessary to collapse cells where all schools participated, as the school non-response adjustment factor was 1.0 regardless of whether cells were collapsed or not. However, such cells were sometimes collapsed to ensure that enough responding students would be available for the student non-response adjustments in a later weighting step. In either of these situations, cells were generally collapsed over the last implicit stratification variable(s) until the violations no longer existed. Within a given country, usually about 10 to 30 final cells were formed after collapsing.

Within the school non-response adjustment group containing school *i*, the non-response adjustment factor was calculated as:

$$f_{1i} = \frac{\sum_{k \in \Omega(i)} w_{1k} enr(k)}{\sum_{k \in \Gamma(i)} w_{1k} enr(k)}$$

where the sum in the denominator was over $\Gamma(i)$, which are the schools, k, within the group (originals and replacements) that participated, while the sum in the numerator was over $\Omega(i)$, which are those same schools, plus the original sample schools that refused and were not replaced. The numerator estimates the population of 15-year-old students in the group, while the denominator gives the size of the population of 15-year-old students directly represented by participating schools. The school non-response adjustment factor ensures that participating schools are weighted to represent all students in the group. If a school did not participate because it had no PISA-D-eligible students enrolled, no adjustment was necessary since this was considered neither non-response nor undercoverage.

Table 8.1 shows the number of school non-response classes that were formed for each country, and the variables that were used to create the cells.

Table 8.1 **Non-response classes**

Country	Number of explicit strata*	Implicit stratification variables	Number of original cells	Number of final cells
Cambodia	25	School Management (2); Shifts (2)	46	13
Ecuador	8	Province (25); Academic calendar (2); ISCED levels (3)	98	20
Guatemala	9	ISCED (3); Modality (4)	23	9
Honduras	9	Gender (5); Department (18)	136	24
Paraguay	18	Region (5)	61	16
Senegal	30	School Management (2); Inspection (59); Gender (3)	103	19
Zambia	20	School Type (4)	45	13

* For details of the explicit stratification, see Table 4.1, in Chapter 4.

Note: ISCED = International Standard Classification of Education.

The within school non-response adjustment

Within each final school non-response adjustment cell, explicit stratum and high/low grade, gender, and school combination, the student non-response adjustment f_{2i} was calculated as:

$$f_{2i} = \frac{\sum_{k \in X(i)} f_{1i} w_{1i} w_{2ik}}{\sum_{k \in \Delta(i)} f_{1i} w_{1i} w_{2ik}}$$

where

$\Delta(i)$ is all assessed students in the final school non-response adjustment cell and explicit stratum-grade-gender-school combination; and

$X(i)$ is all assessed students in the final school non-response adjustment cell and explicit stratum-grade-gender-school combination, plus all others who should have been assessed (i.e., who were absent, but not excluded or ineligible).

The high- and low-grade categories in each country were defined for each to contain a substantial proportion of the PISA-D population in each explicit stratum of larger schools.

The definition was then applied to all schools of the same explicit stratum characteristics

regardless of school size. In most cases, this student non-response factor reduces to the ratio of the number of students who should have been assessed to the number who were assessed. In some cases of small (i.e. fewer than 15 respondents) cell (i.e., final school non-response adjustment cell and explicit stratum-grade-gender-school category combinations) sizes, it was necessary to collapse cells together, and then apply the more complex formula shown above. Additionally, an adjustment factor greater than 2.0 was not allowed for the same reasons noted under school non-response adjustments. If this occurred, the cell with the large adjustment was collapsed with the closest cell within grade and gender combinations in the same school non-response cell and explicit stratum.

Some schools in some countries had extremely low student response levels. In these cases it was determined that the small sample of assessed students within the school was potentially too biased as a representation of the school to be included in the final PISA-D dataset. For any school where the student response rate was below 25%, the school was treated as a non-respondent, and its student data were removed. In schools with between 25% and 50% student response, the student non-response adjustment described above would have resulted in an adjustment factor of between 2.0 and 4.0, so the grade-gender cells of these schools were collapsed with others to create student non-response adjustments.[3]

Trimming the student weights

This final trimming check was used to detect individual student weights that were unusually large compared to those of other students within the same explicit stratum. The sample design was intended to give all students from within the same explicit stratum an equal probability of selection, and therefore equal weight, in the absence of school and student non-response. As already noted, poor prior information about the number of eligible students in each school could lead to substantial violations of this equal weighting principle. Moreover, school and student non-response adjustments, and, occasionally, inappropriate student sampling could, in a few cases, accumulate to give a few students in the data relatively large weights, which adds considerably to the sampling variance. The weights of individual students were therefore reviewed and compared against a threshold of more than five times the median weight of students from the same explicit sampling stratum. Based on this comparison, the trimming of student weights was not required within any country.

The student trimming factor, t_{2ij}, is equal to the ratio of the final student weight to the student weight adjusted for student non-response, and therefore equal to 1.0 for all students. The final weight variable on the data file is the final student weight that incorporates any student-level trimming.

CALCULATING SAMPLING VARIANCE

A replication methodology was employed to estimate the sampling variances of PISA-D parameter estimates. This methodology accounted for the variance in estimates due to the sampling of schools and students. Additional variance due to the use of plausible values from the posterior distributions of scaled scores was captured separately as measurement variance. Computationally, the calculation of these two components could be carried out in a single

program, such as *WesVar 5* (Westat, 2007). The SPSS and SAS macros were also developed as part of the International Association for the Evaluation of Educational Achievement's IDB Analyser.

The balanced repeated replication variance estimator

The approach used for calculating sampling variances for PISA-D estimates is known as balanced repeated replication (BRR), or balanced half-samples; the particular variant known as Fay's method was used. This method is similar in nature to the jackknife method used in other international studies of educational achievement, such as TIMSS and PIRLS, and it is widely documented in the survey sampling literature (see Rust, 1985; Rust and Rao, 1996; Shao, 1996; Wolter, 2007). The major advantage of the BRR over the jackknife method is that the jackknife is not fully appropriate for use with non-differentiable functions of the survey data, most noticeably quantiles, for which it does not provide a statistically consistent estimator of variance. This means that, depending upon the sample design, the variance estimator can be unstable, and despite empirical evidence that it can behave well in a PISA-D-like design, theory is lacking. In contrast the BRR method does not have this theoretical flaw. The standard BRR procedure can become unstable when used to analyse sparse population subgroups, but Fay's method overcomes this difficulty and is well justified in the literature (Judkins, 1990).

The BRR method was implemented for a country where the student sample was selected from a sample of schools, rather than all schools, as follows:

- Schools were paired on the basis of the explicit and implicit stratification and frame ordering used in sampling. The pairs were originally sampled schools, or their replacement in cases of non-participation, except for participating replacement schools that took the place of an original school. For an odd number of schools within a stratum, a triple was formed consisting of the last three schools on the sorted list.
- In certainty schools, variance strata were assigned at the student level using the same procedure described for non-certainty schools. Students were paired on the basis of the ordering used in student sampling.
- Pairs were numbered sequentially, 1 to *H*, with pair number denoted by the subscript *h*. Other studies and the literature refer to such pairs as variance strata or zones, or pseudo-strata.
- Within each variance stratum, one school was randomly numbered as 1, the other as 2 (and the third as 3, in a triple), which defined the variance unit of the school. Subscript *j* refers to this numbering.
- These variance strata and variance units (1, 2, 3) assigned at school level were attached to the data for the sampled students within the corresponding school.
- Let the estimate of a given statistic from the full student sample be denoted as X^*. This was calculated using the full sample weights.
- A set of 80 replicate estimates, X_t^* (where *t* runs from 1 to 80), was created. Each of these replicate estimates was formed by multiplying the survey weights from one of the two schools in each stratum by 1.5, and the weights from the remaining schools by 0.5.

The determination as to which schools received inflated weights, and which received deflated weights, was carried out in a systematic fashion, based on the entries in a Hadamard matrix of order 80. A Hadamard matrix contains entries that are +1 and –1 in value, and has the property that the matrix, multiplied by its transpose, gives the identity matrix of order 80, multiplied by a factor of 80. Details concerning Hadamard matrices are given in Wolter (2007). The choice to use 80 replicates was made at the outset of the PISA project in 2000 and was retained for use in PISA-D. This number was chosen because it is "fully efficient" if the sample size of schools is equal to the minimum number of 150 (in the sense that using a larger number would not improve the precision of variance estimation), and because having too large a number of replicates adds computational burden. In addition the number of replicates must be a multiple of 4.

- In cases where there were three units in a triple, either one of the schools (designated at random) received a factor of 1.7071 for a given replicate, with the other two schools receiving factors of 0.6464; otherwise, the one school received a factor of 0.2929 and the other two schools received factors of 1.3536. The explanation of how these particular factors came to be used is explained in Appendix 12 of the PISA 2000 Technical Report (Adams and Wu, 2002).
- To use a Hadamard matrix of order 80 requires that there be no more than 80 variance strata within a country; otherwise, some combining of variance strata had to be carried out prior to assigning the replication factors via the Hadamard matrix. The combining of variance strata does not cause bias in variance estimation, provided that it is carried out in such a way that the assignment of variance units is independent from one stratum to another within strata that are combined. That is, the assignment of variance units must be completed before the combining of variance strata takes place. This approach was used for PISA-D.
- The reliability of sampling variance estimates for important population subgroups is enhanced if any combining of variance strata that is required is conducted by combining variance strata from different subgroups. Thus in PISA-D, variance strata that were combined were selected from different explicit sampling strata and also, to the extent possible, from different implicit sampling strata.
- The sampling variance estimator is then:

$$V_{BRR}\left(X^*\right) = 0.05\sum_{t=1}^{80}\left\{\left(X_t^* - X^*\right)^2\right\}.$$

The properties of BRR method have been established by demonstrating that it is unbiased and consistent for simple linear estimators (i.e., means from straightforward sample designs), and that it has desirable asymptotic consistency for a wide variety of estimators under complex designs, and through empirical simulation studies.

Reflecting weighting adjustments

This description does not detail one aspect of the implementation of the BRR method. Weights for

a given replicate are obtained by applying the adjustment to the weight components that reflect selection probabilities (the school base weight in most cases), and then recomputing the non-response adjustment replicate by replicate.

Implementing this approach required that the PISA Consortium produce a set of replicate weights in addition to the full sample weight. Eighty such replicate weights were needed for each student in the data file. The school and student non-response adjustments had to be repeated for each set of replicate weights.

To estimate sampling variance correctly, the analyst must use the variance estimation formula above by deriving estimates using the *t*-th set of replicate weights. Because of the weight adjustments (and the presence of occasional triples), this does not mean merely increasing the final full sample weights for half the schools by a factor of 1.5 and decreasing the weights from the remaining schools by a factor of 0.5. Many replicate weights will also be slightly disturbed, beyond these adjustments, as a result of repeating the non-response adjustments separately by replicate.

Formation of variance strata

With the approach described above, all original sampled schools, or their participating replacements, were sorted in stratum order (including refusals, excluded and ineligible schools) and paired. An alternative would have been to pair participating schools only. However, the approach used permits the variance estimator to reflect the impact of non-response adjustments on sampling variance, which the alternative does not. This is unlikely to be a large component of variance in any PISA-D country, but the procedure gives a more accurate estimate of sampling variance.

Notes

1. Note that this is not the same as excluding certain portions of the school population. This also happened in some cases, but cannot be addressed adequately through the use of survey weights.

2. Very small schools with an *ENR* greater than 2 but less than one-half the TCS were also undersampled by a factor of 4 to keep the total number of schools sampled manageable for 4 countries.

3. Chapter 11 describes these schools as being treated as non-respondents for the purpose of response rate calculation even though their student data were used in the analyses.

References

Adams, R. J. and M. Wu (eds.) (2002), *PISA 2000 Technical Report*, OECD Publishing, Paris.

Cochran, W. G. (1977), *Sampling Techniques, 3rd edition*, John Wiley and Sons, New York, NY.

Judkins, D.R. (1990), "Fay's method for variance estimation", *Journal of Statistics*, Vol. 6, pp. 223-229.

Kish, L. (1992), "Weighting for unequal Pi", *Journal of Official Statistics*, Vol. 8/2, pp. 183-200.

Lohr, S. L. (2010), *Sampling: Design and Analysis, Second Edition*, Brooks/Cole, Boston, MA.

Rust, K. (1985), "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics*, Vol. 1/4, pp. 381-397.

Rust, K. and J. N. K. Rao (1996), "Replication methods for analyzing complex survey data", *Statistical Methods in Medical Research: Special Issue on the Analysis of Complex Surveys*, Vol. 5, pp. 283-310

Särndal, C., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York, NY.

Wolter, K. (2007), *Introduction to Variance Estimation, Second Edition*, Springer, New York, NY.