# CHAPTER 8 SURVEY WEIGHTING AND THE CALCULATION OF SAMPLING VARIANCE

Survey weights are required to analyse PISA data, to calculate appropriate estimates of sampling error and to make valid estimates and inferences of the population. The PISA Consortium calculated survey weights for all assessed, ineligible and 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 in each individual participating country.

## SURVEY WEIGHTING

While the students included in the final PISA sample for a given country were chosen randomly, the selection probabilities of the students vary. 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 population.

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

- A school sample design may intentionally over or under-sample certain sectors of the school population: in the former case, so that they could be effectively analysed separately for national purposes, such as a relatively small but politically important province or region, or a sub-population using a particular language of instruction; and in the latter case, 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. 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 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. It is also possible that only part of the PISA-eligible population in a school (such as those 15-year-old students in a particular grade) were represented by its student sample, which also requires weighting to compensate for the missing data from the omitted grades.

[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.

- Student non-response, within participating schools, occurred to varying extents. Sampled students who were PISA-eligible and not excluded, but did not participate in the assessment for reasons such as absences or refusals, will 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 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 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), which were all implemented by the International Association for the Evaluation of Educational Achievement (IEA). The underlying statistical theory for the analysis of survey data 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:

$$W_{ij} = t_{2ij} f_{1i} f_{2ij} f_{1ij}^{A} t_{1i} w_{2ij} w_{1i} \quad \text{MERGEFORMAT}$$

Where:

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

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

$f_{1i}$ 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);

$f_{1ij}^{A}$ is an adjustment factor to compensate for schools in some participating countries where only 15-year-old students who were enrolled in the modal grade for 15-year-old students were included in the assessment;

$f_{2ij}$ 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;

$t_{1i}$ is a school base weight trimming factor, used to reduce unexpectedly large values of $w_{1i}$; 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, this weight is the reciprocal of the selection probability for the school, and is given as:

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

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 school sample size for that stratum.

$MOS_i$ was set as equal to the estimated number of 15-year-old students in the school (denoted as $EST_i$), if it was greater than the predetermined target cluster size (*TCS*), which was 42 students for most countries that did CBA, and 35 for most countries that did CBA minus CPS or PBA. 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;

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

These different values of the MOS are intended to minimize the impact of small schools on the variation of the weights, while recognizing 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 was 150 000 students, with a school sample size of 150, then the sampling interval, $I_1$ = 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 ten 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 $MOS_i$ 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-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 three as the cut-off 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, but to not introduce any substantial bias by altering many student weights to a large degree. Very few school weights were trimmed in any one country, and in most countries no school weights were trimmed.

**The within-school base weight**

The term $w_{2ij}$ is referred to as the within-school base weight. With the PISA 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. This weight is given as:

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

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-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.

In the case of the grade sampling option, for direct sampled grade students, the sampling interval for the extra grade students was the same as that for the PISA students. Therefore, countries with

extra direct-sampled grade students (Iceland) have the same within school student weights for the extra grade students as those for PISA-eligible students from the same school.

Additional weight components were needed for the grade students in Germany and Italy. For these two countries, the extra weight component consisted of the class weight for the selected class(es) (all students were selected into the grade sample in the selected class(es)). In these two countries, the use of whole-classroom sampling for the grade samples resulted in the need for a separate weighting process for the grade samples.

**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 in general 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. Usually, about 10 to 30 such groups were formed within a given country depending upon school distribution with respect to stratification variables. If a country provided no implicit stratification variables, schools were divided into three roughly equal groups, within each explicit stratum, based on their enrolment size. It was desirable to ensure that each group had at least six participating schools, as 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, as they could have caused increased variability in the weights and would have 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. In participating countries with very high overall levels of school non-response after school replacement, the requirement for school non-response adjustment factors to all be below 2.0 was waived.

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

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

where the sum in the denominator is over $\Gamma(i)$, which are the schools, k, within the group (originals and replacements) that participated, while the sum in the numerator is 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-eligible students enrolled, no adjustment was necessary since this was considered neither non-response nor under-coverage.

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**

**The grade non-response adjustment**

Because of perceived administrative inconvenience, individual schools may occasionally agree to participate in PISA but require that participation be restricted to 15-year-old students in the modal grade for 15-year-old students, rather than all 15-year-old students. Since the modal grade generally includes the majority of the population to be covered, such schools may be accepted as participants rather than have the school refuse to participate entirely. For the part of the 15-year-old population in the modal grade, these schools are respondents, while for the rest of the grades in the school with 15-year-old students, such a school is a refusal. To account for this, a special non-response adjustment can be calculated at the school level for students not in the modal grade (and is automatically 1.0 for all students in the modal grade). No countries had this type of non-response for PISA 2015, so the weight adjustment for grade non-response was automatically 1.0 for all students in both the modal and non-modal grades, and therefore did not affect the final weights.

If the weight adjustment for grade non-response had been needed (as it was in earlier cycles of PISA in a few countries), it would have been calculated as follows:

Within the same non-response adjustment groups used for creating school non-response adjustment factors, the grade non-response adjustment factor for all students in school *i*, $f_{1i}^{A}$ , is given as:

$$f_{1i}^{A} = \begin{cases} \sum_{k \in C(i)} w_{1k} enra(k) \Big/ \sum_{k \in B(i)} w_{1k} enra(k) & \text{for students not in the modal grade} \\ 1 & \text{otherwise} \end{cases} \quad \text{MERGEFORMAT}$$

The variable $enra(k)$ is the approximate number of 15-year-old students in school *k* but not in the modal grade. The set B(*i*) is all schools that participated for all eligible grades (from within the non-response adjustment group with school (*i*)), while the set C(*i*) includes these schools and those that only participated for the modal responding grade.

This procedure gives, for each school, a single grade non-response adjustment factor that depends upon its non-response adjustment class. Each individual student has this factor applied to the weight if he/she did not belong to the modal grade, and 1.0 if belonging to the modal grade. In general, this factor is not the same for all students within the same school when a country has some grade non-response.

**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:

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

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 to each contain a substantial proportion of the PISA 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 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, and so the grade-gender cells of these schools were collapsed with others to create student non-response adjustments.[2]

For countries with extra direct grade sampled students (Iceland), care was taken to ensure that student non-response cells were formed separately for PISA students and the extra non-PISA grade students. No procedural changes were needed for Germany and Italy since a separate weighting stream was needed for the grade students.

**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, grade, 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 where the weight was more than four times the median weight of students from the same explicit sampling stratum, it was trimmed to be equal to four times the median weight for that explicit stratum. The trimming of student weights was a rare occurrence, happening in only about 15% of the counties, with only a few cases 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 the great majority of students. The final weight variable on the data file is the final student weight that incorporates any student-level trimming. As in all previous PISA cycles, minimal trimming was required at either the school or the student levels.

**National Option Students**

Other than class-based grade sampling, three countries had national option students, each of which required a separate weighting stream. The weighting stream followed all the usual weighting steps. Mexico had one state which was sampled separately from the Mexico national sample (the state was also covered in the Mexico national sample). Spain had its 17 adjudicated regions in its extra weighting stream (the Spain national sample was a subsample of the adjudicated regional sample, with the addition of schools from the one non-adjudicated region.) The United States had separate weighting streams for each of Puerto Rico, Massachusetts public schools, and North Carolina public schools Massachusetts and North Carolina were also covered in the United States national sample).

[2] 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.

Several other countries also had national option students but in these cases, weighting was done along with the PISA students (Australia, Denmark) if weights were required, or not, if not required (Luxembourg).

### International Options

For both Financial Literacy and the Teacher Questionnaire, no weights were required nor calculated, given the way the samples were selected and the way these data were analysed. The unweighted Financial Literacy response rates were calculated, as were those for the Teacher Questionnaire, to be used as quality indicators, if needed.

## CALCULATING SAMPLING VARIANCE

A replication methodology was employed to estimate the sampling variances of PISA 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 error. 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. For further detail, see *PISA Data Analysis Manual, 2nd edition* (OECD, 2009).

### The balanced repeated replication variance estimator

The approach used for calculating sampling variances for PISA 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 it is well documented in the survey sampling literature (*see* Rust, 1985; Rust and Rao, 1996; Shao, 1996; Wolter, 2007). The major advantage of the BRR method 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-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, 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.

- 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. 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 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, or else 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 & Wu, 2002).

- To use a Hadamard matrix of order 80 requires that there be no more than 80 variance strata within a country, or else that some combining of variance strata 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, and this approach was used for PISA.

- The reliability of 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, variance strata that were combined were selected from different explicit sampling strata and also, to the extent possible, from different implicit sampling strata.

- In some countries, it was not the case that the entire sample was a two-stage design, of first sampling schools and then sampling students within schools. In some countries for part of the sample (and for the entire samples for Cyprus, Iceland, Luxembourg, Macao-China, Malta, Qatar, and Trinidad and Tobago), schools were included with certainty into the sampling, so that only a single stage of student sampling was carried out for this part of the sample. In these cases instead of pairing schools, pairs of individual students were formed from within the same school (and if the school had an odd number of sampled students, a triple of students was formed). The procedure of assigning variance units and replicate weight factors was then conducted at the student level, rather than at the school level.

- In contrast, in one country, the Russian Federation, there was a stage of sampling that preceded the selection of schools. Then the procedure for assigning variance strata, variance units and replicate factors was applied at this higher level of sampling. The schools and students then inherited the assignment from the higher-level unit in which they were located.

- Procedural changes were in general not needed in the formation of variance strata for countries with extra direct grade sampled students (Iceland) since the extra grade sample came from the same schools as the PISA students. However, since all schools in Iceland were certainty schools, students within the schools were paired so that PISA non-grade students were together, PISA grade students were together and non-PISA grade students were together. No procedural changes were required for the grade students for Germany and Italy, since a separate weighting stream was needed in these cases.

- The variance estimator is then:

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

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 re-computing 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 errors 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 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 country, but the procedure gives a more accurate estimate of sampling variance.

**Countries and economies where all students were selected for PISA**

In Iceland, Luxembourg, Macao-China, Malta, and Qatar, all PISA-eligible students were selected for participation in PISA. It might be unexpected that the PISA data should reflect any sampling variance in these countries, but students have been assigned to variance strata and variance units, and the BRR method does provide a positive estimate of sampling variance for two reasons. First, in each country there was some student non-response. Not all PISA-eligible students were assessed, giving sampling variance. Second, the intent is to make inference about educational systems and not particular groups of individual students, so it is appropriate that a part of the sampling variance reflect random variation between student populations, even if they were to be subjected to identical educational experiences. This is consistent with the approach that is generally used whenever survey data are used to try to make direct or indirect inference about some underlying system.