PISA 2022 Technical Report

10 Survey Weighting and the Calculation of Sampling Variance

Survey weights are required to analyse PISA data, to calculate appropriate estimates of population parameters, their 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 population parameters and of standard errors, and to conduct significance tests and create confidence intervals appropriately, taking into account the complex sample design used to select individual student participants for PISA.

Survey weighting

While the students included in the final PISA sample for a given country/economy were chosen randomly, the selection probabilities of the students vary. Survey weights must be incorporated into the analysis to ensure that each participating student appropriately represents the correct number of students in the full PISA population. Sampling weights are used to control the proportional contribution of each participating unit to the overall population estimate.

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

- 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. Note that this is not the same as excluding certain portions of the school population. This also happened in some cases, but this cannot be addressed adequately using survey weights.
- Available information about school size at the time of sampling may not have been completely accurate. If a school had a large student body, the selection probability was based on the assumption that only a sample of students from the school would participate in PISA. But if the school turned out to be smaller than expected, a larger proportion of students would be included. In this scenario, there was a higher probability that the students would be selected in the sample than planned, making their inclusion probabilities higher than those of most other students in the sample. On the other hand, if a school, that was expected to be small, was actually 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.

- 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, 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/economy. 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 may be effective in reducing standard errors (Kish, 1992[1]).
- In countries/economies that opted to participate in the financial literacy study, additional students were selected in all schools. Since the financial literacy sample was also designed to represent the full PISA student population, the weights for the sampled students were adjusted to account for this. Different adjustment factors applied to each student's weight, depending on which assessment form the student was assigned.

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 are used in other international studies of educational achievement such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress of International Literacy study (PIRLS), among others. The underlying statistical theory for the analysis of survey data can be found in Cochran (1977[2]), Lohr (2010[3]) and Särndal, Swensson and Wretman (1992[4]).

Weights are generally 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 four adjustment factors, and can be expressed as:

Formula 10.1

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

Where:

w_{1i} (the school base weight) is calculated as the reciprocal of the probability of inclusion of school i into the sample;

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

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);

w_{2ij} (the within-school base weight) is calculated 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, this weight is the reciprocal of the selection probability for the school, and is calculated as:

Formula 10.2

$$w_{1i} = \begin{Bmatrix} I_g / MOS_i & \text{if } MOS_i < \boldsymbol{I}_g \\ 1 & \text{otherwise} \end{Bmatrix}$$

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.

The measure of size (MOS_i) was set as equal to the estimated number of 15-year-old students in the school (EST_i), if it was greater than the predetermined target cluster size (TCS), which was 42 students for most countries/economies that did a computer-based assessment, and 35 for most countries/economies that did a paper-based assessment. For smaller schools the MOS_i value is given via the following formula, where again, EST_i denotes the estimated number of 15-year-old students in the school:

Formula 10.3

$MOS_i = EST_i$ if $EST_i \geq TCS$;

$= TCS$ if $TCS > EST_i \geq TCS/2$;

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

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

These different values of the measurement of size (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 100 15-year-old students at the time of sample selection then $MOS_i = 100$. And, if the country/economy had a single explicit stratum ($g = 1$) and the total of the MOS_i values of 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 $w_{1i} = 1{,}000/100 = 10$. Thus, the school should represent 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 is larger than the sampling interval. In the case where one or more schools have a MOS_i value that exceeds the relevant sampling interval value (I), these schools become certainty selections, and the value of I is recalculated after removing them.

In the case of replacements, the MOS_i used in the calculation of the school base weight is that of the replacement school (not the original school).

The school base weight trimming factor

Once school base weights were established for each sampled school in the participating country/economy, 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.

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 where the 15-year-old student enrolment exceeded 3 × MAX (TCS, MOS_i) were flagged. For example, if the target cluster size (TCS) was 42 students, then a school flagged for trimming had more than 126 (= 3 x 42) PISA-eligible students, and more than 3 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, ENR_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, with the aim of not introducing any substantial bias by altering many student weights to a large degree. The school trimming happened in 13 participating countries/economies. There were four school weights trimmed for Cambodia and Panama respectively, and six school weights trimmed for Denmark. In the remaining countries/economies where some trimming was needed only one or two school weights were trimmed.

The school non-response adjustment

In order to adjust for the fact that those schools that declined to participate, and were not replaced, were not in general typical of the schools in the sample as a whole, school-level non-response adjustments were made. Within each participating country/economy 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 non-participating schools and their students.

The compositions of the non-response groups varied among countries/economies, but the original adjustment groups for all countries/economies were formed by cross-classifying the explicit and implicit stratification variables used for school sample selection. Usually, about 10 to 40 such groups were formed within a given country/economy depending upon school distribution with respect to stratification variables. If a country/economy 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 groups where all schools participated, as the school non-response adjustment factor was 1.0 regardless of whether groups were collapsed or not. However, since the groups used for school non-response adjustment were also used as the basis for student non-response adjustment, such groups 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, groups were generally collapsed starting from the last implicit stratification variable until the violations no longer existed. In countries/economies with very high overall levels of school non-response after school replacement, explicit strata were sometimes collapsed.

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

Formula 10.4

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

where *enr* (*k*) is the actual enrolment of 15-year-old students in the school at the time of preparation of the student list (and so, in general, is somewhat different from the EST_i), 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 10.1 shows the number of school non-response classes that were formed for each country/economy, and the variables that were used to create the cells.

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:

Formula 10.5

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

where sam_i is the number of students sampled 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 a proportion of students in the school.

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/economies with extra direct-sampled grade students (e.g., 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 France and Germany. The extra weight component consisted of the class weight for 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.

The within-school non-response adjustment

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

Formula 10.6

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

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 participating country/economy were defined so that each grade category contained a substantial proportion of the PISA population in each original explicit stratum or final school non-response adjustment groups where collapsing crossed explicit strata. The definition was then applied to all schools in the same original explicit stratum or in the same final school non-response adjustment group.

In most cases, the 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, adjustments greater than 2.0 were flagged for review, 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 participating countries/economies 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 33%, the school was treated as a non-respondent, and its student data were removed.

For countries/economies with extra PISA immigrant student (Denmark, Finland) or 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 students. No procedural changes were needed for France and Germany 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 original explicit stratum. The sample design was intended to give all students from within the same original 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 relatively large weights, which increases the sampling variance. The student non-response adjusted 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 happened in about 11% of all participating countries/economies.

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 within each explicit stratum, 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

Spain had a financial literacy subsample of its national sample, which required a separate weighting stream. The extra weighting stream followed all the usual weighting steps.

A few other countries/economies also had national option students but, in these cases, weighting was done along with the PISA students (i.e., Denmark, Finland, and Iceland) if weights were required. Specifics about national options are beyond the scope of this report.

International options

For the teacher questionnaire (TQ), special weight factors were applied at the end of weighting in 18 countries/economies to ensure that in the TQ database, the sum of weights of the math and non-math teachers would still approximate the math and non-math teacher population, respectively. For financial literacy, special weight factors were applied at the end of weighting to ensure that in the financial literacy database, the sum of weights of the financial literacy students would still approximate the PISA population. The overall, math, and non-math weighted teacher questionnaire response rates were calculated. The weighted financial literacy response rates were also calculated.

Teacher weighting

While the TQ has been an international option in past cycles, the PISA 2022 cycle is the first cycle in which survey weights were calculated for sampled teachers. This section describes the methodology for calculating teacher weights. Eighteen countries/economies participated in the TQ option. Teachers eligible for TQ were those that were currently teaching the modal grade(s) of PISA-eligible students in the country/economy. In 2022, the TQ option consisted of separate samples of mathematics teachers and 'other' teachers (those not teaching mathematics).

It is possible that a teacher who was identified as a mathematics teacher on the teacher list provided by the school was found to be a non-mathematics teacher based on their response in the TQ, and vice versa. On the rare occasions that this occurred, the teacher weight was calculated based on their classification at the time of selection (i.e., as identified on the teacher list). In the delivery file, the teacher 'type' (mathematics or non-mathematics teacher) identified on the teacher list and the teacher 'type' identified by the teacher in their TQ are both available for analysis purposes.

The TQ weighting methodology followed closely the approach described in the previous section for student weighting. However, there are several differences, and these are described in the subsections that follow.

The TQ school base weight

Because TQ data were collected primarily for use in conjunction with the data of participating PISA students, the set of participating schools identified during student weighting was determined to be the set of participating schools for teacher weighting. Therefore, any responding teachers outside of these schools were dropped from the TQ sample. The final school weights from student sampling were used for calculating TQ weights. These final school weights incorporate school base weight trimming and school non-response adjustments, and these are described in some detail earlier in this chapter.

It is possible that a participating school did not have a teacher list completed and, as a result, had no teachers sampled. Such schools will usually be ineligible for the TQ, because they would have no PISA-eligible students in the modal grade. However, it is possible that a TQ-eligible school had no teachers listed or sampled. For such schools, an additional school non-response stage was carried out. There were five countries for which this extra adjustment was required, with the number of schools shown in parentheses – Australia (3), Brazil (5), Colombia (2), Hong Kong (5), and Panama (5). These schools were coded as nonrespondents for the purpose of TQ weighting, and the final school weights of other participating schools in the same final school non-response adjustment cell from student weighting were increased to account for this additional school non-response.

Where the teacher response rate within a participating school was low (or 0%), this was handled through teacher non-response adjustment. A school-level teacher participation rate was calculated and included as a variable on the teacher delivery file. This information can be used as a measure to provide data users with information about the quality of school-level TQ data.

The within-school teacher base weight

The within-school teacher base weight was calculated in the same way as the within-school student base weight. Since the samples of mathematics teachers and non-mathematics teachers are selected independently, teacher weights for mathematics teachers within a particular school will differ from weights for non-mathematics teachers. However, within a particular school, all mathematics teachers have the same within-school base weight, and all non-mathematics teachers have the same within-school base weight. The formula for within-school teacher base weights can be written as follows:

Formula 10.7

$$w_{2ikl} = enr_{ik}/sam_{ik}$$

where k=1 or 2, to indicate mathematics or non-mathematics teachers, enr_{i1} and sam_{i1} are the number of mathematics teachers and *sampled* mathematics teachers respectively in school *i*, and enr_{i2} and sam_{i2} are the number of non-mathematics teachers and *sampled* non-mathematics teachers respectively in school *i*.

The within-school teacher non-response adjustment

The teacher non-response adjustment followed the same approach as the student non-response adjustment. For teachers, the only information available besides the final school non-response cell and explicit stratum is the teacher type (mathematics or non-mathematics teacher). Within each final school non-response adjustment cell, explicit stratum, and teacher type, and school combination, the teacher non-response adjustment f_{2i} was calculated as:

Formula 10.8

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

where,

$\Delta(i)$ is all participating teachers in the final school non-response adjustment cell and explicit stratum-teacher type-school combination; and,

$X(i)$ is all participating and non-participating teachers in the final school non-response adjustment cell and explicit stratum-teacher type-school combination. Ineligible teachers are excluded from the calculation. Note that there no *excluded* teachers.

Collapsing of teacher non-response adjustment cells was done as needed to ensure at least 15 participating teachers were in each final adjustment cell. Because the number of sampled mathematics teachers in each school was often small, collapsing across schools was always required for mathematics teachers, and there were instances where it was necessary to collapse teacher types.

Trimming the teacher weights

The PISA sample design is intended to produce a self-weighting sample of *students*, in the absence of school and student non-response. There are several reasons why final student weights vary, and these are described at the beginning of this chapter. However, extreme outlier student weights are typically due to poor frame data on school-level student enrolment. As described in the student weight trimming section, extreme student weights are trimmed in order to reduce the sampling variance.

In contrast, the PISA sample design was not intended to produce self-weighting samples of teachers. Schools were sampled proportional to student enrolment, and while the number of mathematics and non-mathematics teachers in a school can be expected to be correlated with student enrolment, this relationship varies from school to school, and no steps were taken to reduce the weight variability of the teacher samples. Since teacher weights vary considerably *by design*, there was no clear basis to identify 'outlier' teacher weights. It was decided that no trimming of teacher weights would be carried out.

Calculating sampling variance

A replication methodology is employed to estimate the sampling variances of the PISA parameter estimates. This methodology accounts for the variance in estimates due to the sampling of schools and students. Additional variance due to the use of plausible values drawn from the posterior distributions of scaled scores is captured separately as measurement or imputation error. Computationally the calculation of these two components could be carried out using a single program, such as *WesVar 5*, or with the IDB Analyzer using R, SPSS and SAS macros developed for this purpose.

The balanced repeated replication variance estimator

The specific replication 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 PIRLS, and it is well documented in the survey sampling literature [see Rust (1985[5]); Rust and Rao (1996[6]); Rao and Shao (1996[7]); Wolter (2007[8])]. The major advantage of the balanced repeated replication (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, and for which the jackknife methods 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 literature (Judkins, 1990[9]).

For a country/economy where the student sample was selected from a sample of schools, rather than all schools, the BRR method was implemented 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. In the case of an odd number of schools within a stratum, a triplet 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, variance 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 triplet), 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 2 schools in each stratum by 1.5, and the weights from the remaining school 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 the same order. Details concerning Hadamard matrices are given in Wolter (2007[8]). 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 3 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 2 schools receiving factors of 0.6464, or else the one school received a factor of 0.2929 and the other 2 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, 2022[10]).
- To use a Hadamard matrix of order 80 requires that there be no more than 80 variance strata within a country/economy, 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/economies, 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/economies for part of the sample (and for the entire samples for Brunei, Iceland, Macao (China), Malta, North Macedonia, and Qatar), 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, there could have been a stage of sampling that precedes the selection of schools. Then the procedure for assigning variance strata, variance units and replicate factors would be applied at this higher level of sampling. The schools and students would then inherit the assignment from the higher-level unit in which they were located. No countries/economies used such a three-stage design for PISA 2022.
- Procedural changes were in general not needed in the formation of variance strata for countries/economies 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 France and Germany, since a separate weighting stream was needed in these cases.

The variance estimator for the BRR method is then calculated using the following formula:

Formula 10.7

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

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

Implementing this approach required that the PISA Consortium produce a set of replicate weights in addition to the full sample weight. 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 trimming the school base weight, re-computing the school non-response adjustment for each replicate, applying the adjustment for student selection (the student base weight component), computing the student non-response adjustment for the replicate, and trimming the student non-response adjusted weight. The school and student non-response adjustments were recalculated and applied to each set of replicate weights using the methodology described earlier in this chapter. Like the full-sample adjusted student weight, the replicate adjusted student weights are provided as variables in the data file.

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/economy, but the procedure gives a more accurate estimate of sampling variance.

Countries/economies where all students were selected for PISA

In Brunei, Iceland, Macao (China), and Malta, 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/economies, but students have been assigned to variance strata and variance units, and the balanced repeated replication (BRR) method does provide a positive estimate of sampling variance for two reasons. First, in each country/economy there was some student non-response. Not all PISA-eligible students were assessed, resulting in 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 of the 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.

Variance estimation for the TQ sample

The TQ sample used the same variance estimation approach as the student sample. Since the participating schools for the student sample were used as the participating schools for the teacher sample, the full sample final school weight for the student sample was also the full sample final school weight for the teacher sample. Similarly, the replicate school weights for the student sample were used as the replicate school weights for the teacher sample. For certainty schools, instead of pairing schools, pairs of individual teachers were formed from within the same school and the procedure of assigning variance units and replicate weight factors was then conducted at the teacher level, rather than at the school level. Teachers were sorted by teacher type before pairing was done, to maximise the chance of pairing teachers of the same teacher type.

References

Adams, R. and M. Wu (eds.) (2022), *Programme for International Student Assessment (PISA): PISA 2000 Technical Report*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264199521-en. [10]

Cochran, W. (1977), *Sampling Techniques (3rd ed.)*, John Wiley and Sons. [2]

Judkins, D. (1990), “Fay’s method for variance estimation”, *Journal of Official Statistics*, Vol. 6, pp. 223-229. [9]

Kish, L. (1992), “Weighting for unequal Pi”, *Journal of Official Statistics*, Vol. 8/2, pp. 183-200. [1]

Lohr, S. (2010), *Sampling: Design and Analysis*, Brooks/Cole, Boston, MA. [3]

Rao, J. and J. Shao (1996), "On balanced half-sample variance estimation in stratified random sampling", *Journal of the American Statistical Association*, Vol. 91, pp. 343-348. [7]

Rust, K. (1985), "Variance estimation for complex estimators in sample surveys", *Journal of Official Statistics*, Vol. 1/4, pp. 381-397. [5]

Rust, K. and J. 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. [6]

Särndal, C., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York, NY. [4]

Wolter, K. (2007), *Introduction to Variance Estimation*, Springer, New York, NY. [8]

Chapter 10 tables

Table 10.1. School Non-response classes

Country/Economy	Number of explicit strata*	Implicit stratification variables	Number of original cells	Number of final cells
Albania	12	ISCED level (3), Gender (5)	70	16
Argentina	21	Department (19); Location (2); Level (8); Performance (5)	193	43
Australia	25	Geographic Location (3); School gender composition (3); School socio-economic Level (11); ISCED level (3)	385	77
Austria	18	Region (9); Percentage of girls (5); Programme for Statut schools (3)	276	29
Baku (Azerbaijan)	5	None	15	11
Belgium	31	School type – French Community (4), German and Flemish Community (1); Grade repetition – Flemish and French Community (5), German Community and some Flemish and French Community (1); Percentage of Girls – Flemish and French Community (4), German Community and some Flemish and French Community (1)	164	32
Brazil	20	State (27); ISCED level (5); Urbanisation (2); Capital/Country (2); IDH Quintiles (5); School gender composition (3)	506	63
Brunei	8	Sixth Form (3); District (4)	17	5
Bulgaria	3	Type of school (3)	9	9
Cambodia	18	School management (2); Shifts (2)	40	23
Canada	67	Urbanicity (2); Funding (2); ISCED Level (3)	208	38
Chile	14	School Type (4); National test score level (4); Percentage of girls (6); Urbanicity (2); Geographic zone (4)	177	28
Chinese Taipei	19	Funding (2); Region (6); School gender composition (3); Municipality (2); Shift offerings (2)	141	35
Colombia	6	Regional entities (96); Main shift (2); School gender composition (5)	176	33
Costa Rica	6	Zone (2); Track (2); Shift (2); Education regions (27); ISCED level (3)	112	34
Croatia	7	Region (6); School gender composition (3)	56	20
Cyprus	8	Urbanisation (2); Language (2)	16	9
Czechia	32	Region for school types 3, 4, 5 (14); Gender (3)	146	37
Denmark	6	School type (7); ISCED level (3); Urbanisation (5); Region (5); FO group (3)	152	42
Dominican Republic	10	Shift (6); School size (4); Programme (4)	88	23
El Salvador	28	Founding (2); ISCED level (3); Study Commitment (3)	107	26
Estonia	4	School type (3); Urbanicity (2); County (15); Funding (2)	71	15
Finland	30	Immigrant cluster (6); Regional State Administrative agencies (7); School type (5)	62	24
France	22	Secteur (2)	32	14
Georgia	9	Language (9)	22	9
Germany	18	State for SEN and vocational schools only (16); School type for Normal schools (6)	68	24
Greece	3	Funding and region (15); School type (4)	100	26
Guatemala	8	ISCED (2); Modality of teaching (4)	25	11
Hong Kong (China)	5	Student academic intake (4); School gender composition (3)	21	8
Hungary	6	Geographical region of Hungary (7); Average mathematics performance in the National ABC 2020 (6)	132	49
Iceland	24	Urbanicity (2)	23	10
Indonesia	4	School type (5); Funding (2); Region (8)	95	48
Ireland	9	School Gender Composition (4); Socio-Economic Status Quartile (4)	73	21

Country/Economy	Number of explicit strata*	Implicit stratification variables	Number of original cells	Number of final cells
Israel	13	ISCED level (3); Group size (2); Socio-Economic Status (3); Geographic/Administrative District (2)	71	26
Italy	36	Region (20); Types of School (2)	107	32
Jamaica	15	Gender (3); School types (5)	41	15
Japan	4	Levels of proportion of students taking University/College Entrance Exams (4)	14	9
Jordan	8	Region (3); Urbanisation (2); School gender composition (3); Level (2); Shift (2)	96	36
Kazakhstan	19	ISCED level (2); Location (2); Language (3); Funding (2); Shifts (2)	190	69
Korea	6	Urbanisation (3); School gender composition (3)	26	15
Kosovo	8	Urbanisation (2); ISCED (3)	26	11
Latvia	4	School type (4)	15	11
Lithuania	21	School language 2 (4); School location (5); School type (5); School type 2 (2)	45	18
Macao (China)	10	Gender (3); School orientation (2)	18	9
Malaysia	10	School type (18); Location (2); Gender (3); ISCED level (2)	32	11
Malta	3	N/A	9	7
Mexico	12	School program (8); Urbanisation (2)	45	15
Mongolia	16	Property type (3); ISCED orientation (2); ISCED level (3)	27	14
Montenegro	12	Gender (3)	19	15
Morocco	12	Milieu (2); Type (2)	31	22
Netherlands	10	N/A	28	10
New Zealand	4	School decile (4); School authority (2); School gender composition (3); Urbanicity (2)	41	14
North Macedonia	9	Urbanisation (2)	14	9
Norway	2	None	6	3
Palestinian Authority	7	Region (2); Gender (3); District (25)	121	35
Panama	16	Educational region (16); ISCED level (3); Programme orientation (4); Language of test (3)	98	18
Paraguay	19	Region (5)	66	20
Peru	4	Region (26); School gender composition (3); School type (4)	107	30
Philippines	16	School Management (2); Type of Community (3); ISCED Level (3); Gender Composition (5)	73	24
Poland	4	Private/Public (2); Locality size (4); School gender composition (3)	41	6
Portugal	26	ISCED (3); School management (2); School Location (3); Curriculum (3)	97	35
Qatar	4	Level (5); School gender composition (3); Language (2); Programme orientation (3)	39	13
Republic of Moldova	28	Funding (2); ISCED program orientation (6)	38	14
Romania	6	School location area (2); Development regions (8)	46	18
Saudi Arabia	30	Education District (47); School Level (2)	104	37
Serbia	22	Region implicit (5); School type implicit (7); Language (2)	45	25
Singapore	4	Gender (3)	5	4
Slovak Republic	24	T9 - Three-year average of scores in national testing in math and Slovak (Hungarian) language (7); School type (6); Language (3); Funding (3)	146	32
Slovenia	7	Location (5); School Gender Composition (3)	149	33
Spain	40	Linguistic model – for Basque Country only (3), other regions (1)	121	100
Sweden	8	Geographic LAN for upper secondary only (21); Responsible authority, if upper secondary (3); Percentage of immigrant students (3); Income quartiles, if ISCED2 (4)	65	21
Switzerland	15	Sponsorship (2); School type (33); Canton (30); Foreign Speaking Student Share (3)	197	32
Thailand	15	Public/Private (2); Region (9); Urbanisation (2); School gender composition (3)	135	33
Türkiye	36	Statistical Region Unit (12); Location (2); Gender (3)	191	27

Country/Economy	Number of explicit strata*	Implicit stratification variables	Number of original cells	Number of final cells
Ukraine (18 of 27 Regions)	49	ISCED Orientation (3); Language (3)	87	23
United Arab Emirates	47	School gender (3); Language (3); ISCED (3); Programme (2)	146	73
United Kingdom (excl. Scotland)	34	Gender (3); School performance – England (6) and Wales (5); Local authority (7)	332	47
United Kingdom (Scotland)	8	Gender (3); Area type (6)	32	13
United States of America	8	Grade span (5); Urbanisation (4); Minority status (2); Gender (3); State (51)	210	20
Uruguay	11	Location/Urbanisation (4); School gender composition (4)	40	16
Uzbekistan	27	Specialization (2)	49	19
Viet Nam	15	Region (6); Province (63); School type (4); Study commitment (2)	157	29