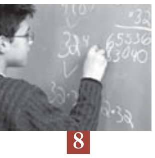
8

Survey Weighting and the Calculation of Sampling Variance



Survey weights were required to analyse PISA 2003 data, to calculate appropriate estimates of sampling error, and to make valid estimates and inferences. The 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, to conduct significance tests and to create confidence intervals appropriately, given the sample design for PISA in each individual country.

SURVEY WEIGHTING

Students included in the final PISA sample for a given country are not all equally representative of the entire student population, despite random sampling of schools and students for selecting the sample. Survey weights must therefore be incorporated into the analysis.

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,[1] such as very small or geographically remote schools.
- Information about school size available at the time of sampling may not have been completely accurate. If a school was expected to be very large, the selection probability was based on the assumption that only a sample of its students would be selected for PISA. But if the school turned out to be quite small, all students would have to be included and would have, overall, a higher probability of selection in the sample than planned, making these inclusion probabilities higher than those of most other students in the sample. Conversely, if a school thought to be small turned out to be large, the students included in the sample would have had 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 eligible population in a school (such as those 15-year-olds in a single 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. Students of the kind that could not be given achievement test scores (but were not excluded for linguistic or disability reasons) will be under-represented in the data unless weighting adjustments are made.
- Trimming 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. This can lead to unstable estimates – large sampling errors – but cannot be estimated well. Trimming weights introduces a small bias into estimates, but greatly reduces standard errors.

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: the Third International Mathematics and Science Study (TIMSS), the Third International Mathematics and Science Study–Repeat (TIMSS-R), the Civic Education Study (CIVED), and the Progress in International Reading Literacy Study 2001 (PIRLS), which were all implemented by the International Association for the Evaluation of

Educational Achievement (IEA), and also in the International Assessment of Educational Progress (IAEP, 1991). (See Cochran, 1977 and Särndal *et al.,* 1992, for the underlying statistical theory on survey sampling texts.)

The weight, W_{ij}, for student *j* in school *i* consists of two base weights – the school and the within-school – and five adjustment factors, and can be expressed as:

$$W_{ij} = t_{2ij} f_{1i} f_{1ij}^{A} t_{1i} w_{2ij} w_{1i} \quad (8.1)$$

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 the fact that, in some countries, in some schools only 15-year-old students who were enrolled in the modal grade for 15-year-olds were included in the assessment;
- t_{1i} is a school trimming factor, used to reduce unexpectedly large values of w_{1i}; and
- t_{2ij}, is a student 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 probability proportional-to-size school sampling method used in PISA, this is given as:

$$w_{1i} = \begin{cases} int\left(g/i\right) \big/ mos\left(i\right) & \text{if } mos\left(i\right) < int\left(g/i\right) \\ 1 & \text{otherwise} \end{cases} \quad (8.2)$$

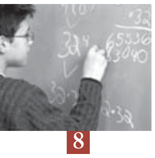
The term $mos\left(i\right)$ denotes the measure of size given to each school on the sampling frame.

Despite country variations, $mos\left(i\right)$ was usually equal to the estimated number of 15-year-olds in the school, if it was greater than the predetermined target cluster size (35 in most countries).

If the enrolment of 15-year-olds was less than the Target Cluster Size (TCS), then $mos\left(i\right) = TCS$.

The term $int\left(g/i\right)$ denotes the sampling interval used within the explicit sampling stratum *g* that contains school *i* and is calculated as the total of $mos\left(i\right)$ values for all schools in stratum *g*, divided by the school sample size for that stratum.

Thus, if school *i* was estimated to have 100 15-year-olds at the time of sample selection, $mos\left(i\right) = 100$. If the country had a single explicit stratum (*g*=1) and the total of the values over all schools was 150 000, with a school sample size of 150, then $\text{int}\left(1/i\right) = 150000/150 = 1000$, for school *i* (and others in the sample), giving $w_{1i} = 1000/100 = 10.0$. Roughly speaking, the school can be thought of as representing about



10 schools from 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 $w_{1i} = 1$.

The school 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 see if the school weights required trimming. The school trimming factor t_{1i}, is the ratio of the trimmed to the untrimmed school base weight, and is equal to 1.0000 for most schools and therefore most students, and never exceeds this value. (See Table 8.1 for the number of school records in each country that received some kind of base weight trimming.)

The school-level trimming adjustment was applied to schools that turned out to be much larger than was believed at the time of sampling – where 15-year-old enrolment exceeded $3 \times \max(TCS,\ mos(i))$. For example, if $TCS = 35$, then a school flagged for trimming had more than 105 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 $3 \times \max(TCS,\ mos(i))$ in the school base weight formula.

The student base weight

The term w_{2ij} is referred to as the student base weight, which with the PISA procedure for sampling students, did not vary across students (j) within a particular school i. This is given as:

$$w_{2ij} = {}^{enr(i)}\!\big/_{sam(i)} \qquad (8.3)$$

where $enr(i)$ is the actual enrolment of 15-year-olds in the school (and so, in general, is somewhat different from the estimated $mos(i)$), and $sam(i)$ is the sample size within school i. It follows that if all students from the school were selected, then $w_{2ij} = 1$ for all eligible students in the school. For all other cases $w_{2ij} > 1$.

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. Several groups of somewhat similar schools were formed within a country, and 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 15 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 stratum, based on their size (small, medium or large). It was desirable to ensure that each group had at least six participating schools, as small groups can lead to unstable weight adjustments, which in turn would inflate the sampling variances. However, 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. Adjustments

greater than 2.0 were flagged for review, as they can cause increased variability in the weights, and lead to an increase in sampling variances. In either of these situations, cells were generally collapsed over the last implicit stratification variable(s) until the violations no longer existed. In countries with very high overall levels of school non-response after school replacement, the requirement for school non-response adjustment factors all to be below 2.0 was waived.

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

$$f_{1i} = \frac{\sum\limits_{k\in\Omega(i)} w_{1k}\, enr\left(k\right)}{\sum\limits_{k\in\Gamma(i)} w_{1k}\, enr\left(k\right)} \qquad (8.4)$$

where the sum in the denominator is over $\Gamma(i)$, the schools within the group (originals and replacements) that participated, while the sum in the numerator is over $\Omega(i)$, those same schools, plus the original sample schools that refused and were not replaced. The numerator estimates the population of 15-year-olds in the group, while the denominator gives the size of the population of 15-year-olds 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 eligible students enrolled, no adjustment was necessary since this was 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.

Grade non-response adjustment

In two countries (Denmark and the United States), several schools agreed to participate in PISA, but required that participation be restricted to 15-year-olds in the modal grade for 15-year-olds, rather than all 15-year-olds, because of perceived administrative inconvenience. Since the modal grade generally included the majority of the population to be covered, some of these schools were accepted as participants. For the part of the 15-year-old population in the modal grade, these schools were respondents, while for the rest of the grades in the school with 15-year-olds, this school was a refusal. This situation occasionally arose for a grade other than the modal grade because of other reasons, such as other testing being carried out for certain grades at the same time as the PISA assessment. To account for this, a special non-response adjustment was calculated at the school level for students not in the modal grade (and was automatically 1.0 for all students in the modal grade).

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:

$$f_{1i}^{A} = \begin{cases} \sum\limits_{k\in C(i)} w_{1k}\, enra\left(k\right) \Big/ \sum\limits_{k\in \mathrm{B}(i)} w_{1k}\, enra\left(k\right) & \text{for students not in the modal grade} \\ 1 & \text{otherwise} \end{cases} \qquad (8.5)$$

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.

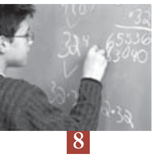


Table 8.1. ■ **Non-response classes**

	Implicit stratification variables used to create school non-response cells (within explicit stratum), and number of original and final cells	**Number of original cells**	**Number of final cells**
Australia	Urban/rural (2)	46	30
Austria	Size (large/small)	30	28
Belgium	Flanders – school proportion of overage students (continuous); French Community – school size (3), school proportion of overage students (continuous); German Community – school type (3), school size (4)	222	46
Brazil	School type (3), urban/rural (2), index of school infrastructure (4)	51	38
Canada	Public/private (2), urban/rural (2)	165	71
Czech Republic	Regions (14) for four school types	140	135
Denmark	School type (4), county (15)	44	18
Finland	Size (3)	35	35
France	Size (3)	18	10
Germany	School type (5) for normal school, for state (16), for vocational schools	67	37
Greece	School type (4), public/private	30	13
Hong Kong-China	For strata 1 and 2, academic intake (3), for independent schools (stratum 3) local or international funding (2)	8	7
Hungary	geographic region (7+1 for missing) for strata 1-4, for stratum 5, TIMSS explicit (TIMSS population variable with two levels) and implicit (20 regions and three levels of urbanization) stratifiers	87	43
Iceland	Urban/rural, school size (4)	33	30
Indonesia	School type (5), public/private (2), national achievement score categories (3)	202	190
Ireland	School type (3), school gender composition categories (5)	24	13
Italy	Public/private (2)	74	30
Japan	Levels (4) of proportions of students taking university or college entrance exams	15	13
Korea	School level (2)	11	10
Latvia	Urbanicity (3), school type (3)	20	8
Liechtenstein	None, three cells formed based on sizes	3	3
Luxembourg	Size (3)	10	4
Macao-China	Size classes (3) for strata 2 and 3	7	7
Mexico	School type (6), urban/rural (2), school level (3), program (3 or 4 depending on school level)	299	259
Netherlands	School type (6)	10	6
New Zealand	Public/private (2), socio-economic status category (3), urban/rural (2)	11	9
Norway	Size (3)	12	7
Poland	Urbanicity (4)	7	5
Portugal	Public/private (2), socio-economic status category (4)	28	20
Russian Federation	School type (3), urbanicity (5) [no school non-response adjustments]	169	157
Serbia	Urban/rural, school type (7), Hungarian students or not	68	64
Slovak Republic	School type (9), language (2), authority (9)	89	53
Spain	For Catalonia: size of town (3), province (numerous); for other regions: province (numerous)	107	107
Sweden	School level (2), income quartile (4), responsible authority(2), urbanicity (5), geographic area (many) –various combinations of these depending on explicit stratum	45	20
Switzerland	School type (many levels), canton (many levels)	171	84
Thailand	Region (13)	58	39
Tunisia	Levels of grade repetition for three school levels (numerous)	41	14
Turkey	School type (18)	123	112
United Kingdom	England – school type (3), exam grade (7), gender (3), region (4, derived from 150 levels of LEA); Wales – secondary/independent, exam grade (4) for secondary schools; Northern Ireland – school type (3), exam grade bands (7), region (5); Scotland – school size (3).	116	47
United States	Gradprop (5), public/private (2), region (4), urbanicity (8), minstat (2)	172	39
Uruguay	Program type (3-7 levels depending on explicit stratum), shift (4 or 5 depending on program, for several strata and are for another stratum), area (3) for one stratum	63	45

This procedure gave, for each school, a single grade non-response adjustment factor that depended upon its non-response adjustment class. Each individual student received this factor value if they did not belong to the modal grade, and 1.0000 if they belonged to the modal grade. In general, this factor is not the same for all students within the same school.

Student non-response adjustment

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

$$f_{2i} = \frac{\sum_{k \in \mathrm{X}(i)} f_{1i} w_{1i} w_{2ik}}{\sum_{k \in \Delta(i)} f_{1i} w_{1i} w_{2ik}} \qquad (8.6)$$

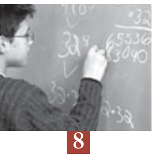
where the set $\Delta(i)$ is all assessed students in the school / grade combination and the set $X(i)$ is all assessed students in the school / grade combination plus all others who should have been assessed (*i.e.* who were absent, but not excluded or ineligible). The high and low grade categories in each country were defined so as to each contain a substantial proportion of the PISA population.

In most cases, this student non-response factor reduces the ratio of the number of students who should have been assessed to the number who were assessed. In some cases of small cells (*i.e.* school/grade category combinations) sizes (fewer than ten respondents), it was necessary to collapse cells together, and then the more complex formula above applied. 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 in the same school non-response cell.

Some schools in some countries had very low student response levels. In these cases it was determined that the small sample of assessed students was potentially too biased as a representation of the school to be included in the PISA data. For any school where the student response rate was below 25 per cent, the school was therefore treated as a non-respondent, and its student data were removed. In schools with between 25 and 50 per cent student response, the student non-response adjustment described above would have resulted in an adjustment factor of between 2.0000 and 4.0000, and so these schools were collapsed with others to create student non-response adjustments.[2]

Trimming student weights

This final trimming check was used to detect student records 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 principle. Moreover, school, grade and student non-response adjustments, as well as, 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 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 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 1.0000 for the great majority of students. The final weight variable on the data file was called *w_fstuwt*, which is the final student weight that incorporates any student-level trimming. Table 8.2 shows the number of students with weights trimmed at this point in the process (*i.e.* $t_{2ij} < 1.0000$) for each country and the number of schools for which the school base weight was trimmed (*i.e.* $t_{1i} < 1.0000$).

Table 8.2 ■ School and student trimming

Country	Number of schools trimmed	Number of students trimmed
Australia	1	0
Austria	0	0
Belgium	0	0
Belgium-Flanders	0	0
Belgium-French	0	0
Belgium-German	0	0
Brazil	0	0
Canada	0	0
Czech Republic	0	0
Denmark	0	0
Finland	0	0
France	0	0
Germany	0	0
Greece	0	0
Hong Kong-China	1	0
Hungary	0	6
Iceland	0	0
Indonesia	5	0
Ireland	0	0
Italy	0	0
Japan	0	0
Korea	0	0
Latvia	0	0
Liechtenstein	0	0
Luxembourg	0	0
Macao-China	0	35
Mexico	0	107
Netherlands	5	0
New Zealand	0	0
Norway	0	0
Poland	0	0
Portugal	1	0
Russian Federation	11	0
Serbia	0	0
Slovak Republic	0	0
Spain	0	0
Sweden	1	0
Switzerland	0	91
Thailand	0	0
Tunisia	0	0
Turkey	1	0
United Kingdom	2	0
England	1	0
Northern Ireland	1	0
Wales	0	0
Scotland	0	0
United States	2	0
Uruguay	0	0

CALCULATING SAMPLING VARIANCE

To estimate the sampling variances of PISA estimates, a replication methodology was employed. This reflected 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, although computationally the two components can be carried out in a single program, such as WesVar 4 (Westat, 2000).

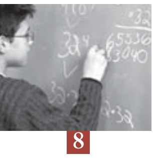
The balanced repeated replication variance estimator

The approach used for calculating sampling variances for PISA is known as Balanced Repeated Replication (BRR), or Balanced Half-Samples; the particular variant known as Fay's method was used. This method is very similar in nature to the Jackknife method used in previous international studies of educational achievement, such as TIMSS, and it is well documented in the survey sampling literature (Rust, 1985; Rust and Rao, 1996; Shao, 1996; Wolter, 1985). The major advantage of BRR over the Jackknife is that the Jackknife method is not fully appropriate for use with non-differentiable functions of the survey data, most noticeably quantiles. It provides unbiased estimates, but not consistent ones. This means that, depending upon the sample design, the variance estimator can be very unstable, and despite empirical evidence that it can behave

well in a PISA-like design, theory is lacking. In contrast, BRR does not have this theoretical flaw. The standard BRR procedure can become unstable when used to analyse sparse population subgroups, but Fay's modification overcomes this difficulty, and is well justified in the literature (Judkins, 1990).

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

- 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 pairs that included a participating replacement if an original refused. For an odd number of schools within a stratum, a triple was formed consisting of the last school and the pair preceding it.
- 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 (the primary sampling unit, PSU) 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 are 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 is 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 sampling weights from one of the two PSUs in each stratum by 1.5, and the weights from the remaining PSUs by 0.5. The determination as to which PSUs 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. (Examples of Hadamard matrices are given in Wolter, 1985.)
- 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* (OECD, 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 any 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.
- 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, to the extent possible, from different implicit sampling strata also.



- 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. In some countries for part of the sample (and for the entire samples for Iceland, Macao-China, Liechtenstein and Luxembourg), 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 also). 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 a few countries there was a stage of sampling that preceded the selection of schools, for at least part of the sample. This was done in a major way in the Russian Federation and Turkey. In these cases there was a stage of sampling that took place before the schools were selected. 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.
- The variance estimator is then:

$$V_{BRR}\left(X^*\right)=0.05\sum_{t=1}^{80}\left\{\left(X_t^*-X^*\right)^2\right\} \qquad (8.7)$$

The properties of BRR 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 glosses over 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 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 instead of the full sample weight. 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, by contrast to other international education assessments such TIMSS and TIMSS-R that have paired participating schools only. However, these studies did not use an approach reflecting the impact of non-response adjustments on sampling variance. This is unlikely to be a big component of variance in any PISA country, but the procedure gives a more accurate estimate of sampling variance.

Countries where all students were selected for PISA

In Iceland, Liechtenstein and Luxembourg, all eligible students were selected for PISA. It might be considered surprising 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 formula does give a positive estimate of sampling variance for three reasons. First, in each country there was some student non-response, and, in the case of Iceland and Luxembourg, some school non-response. Not all eligible students were assessed, giving sampling variance. Second, only 55 per cent of the students were assessed in reading and science. Third, the issue 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.

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] Chapter 12 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.

READER'S GUIDE

Country codes

The following country codes are used in this report:

OECD countries

AUS	Australia
AUT	Austria
BEL	Belgium
BEF	Belgium (French Community)
BEN	Belgium (Flemish Community)
CAN	Canada
CAE	Canada (English Community)
CAF	Canada (French Community)
CZE	Czech Republic
DNK	Denmark
FIN	Finland
FRA	France
DEU	Germany
GRC	Greece
HUN	Hungary
ISL	Iceland
IRL	Ireland
ITA	Italy
JPN	Japan
KOR	Korea
LUX	Luxembourg
LXF	Luxembourg (French Community)
LXG	Luxembourg (German Community)
MEX	Mexico
NLD	Netherlands
NZL	New Zealand
NOR	Norway
POL	Poland
PRT	Portugal
SVK	Slovak Republic
ESP	Spain
ESB	Spain (Basque Community)
ESC	Spain (Catalonian Community)
ESS	Spain (Castillian Community)
SWE	Sweden
CHE	Switzerland
CHF	Switzerland (French Community)
CHG	Switzerland (German Community)
CHI	Switzerland (Italian Community)
TUR	Turkey
GBR	United Kingdom
IRL	Ireland
SCO	Scotland
USA	United States

Partner countries

BRA	Brazil
HKG	Hong Kong-China
IND	Indonesia
LVA	Latvia
LVL	Latvia (Latvian Community)
LVR	Latvia (Russian Community)
LIE	Liechtenstein
MAC	Macao-China
RUS	Russian Federation
YUG	Serbia and Montenegro (Serbia)
THA	Thailand
TUN	Tunisia
URY	Uruguay

List of abbreviations

The following abbreviations are used in this report:

ACER	Australian Council for Educational Research
AGFI	Adjusted Goodness-of-Fit Index
BRR	Balanced Repeated Replication
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CITO	National Institute for Educational Measurement, The Netherlands
CIVED	Civic Education Study
DIF	Differential Item Functioning
ESCS	Economic, Social and Cultural Status
ENR	Enrolment of 15-year-olds
ETS	Educational Testing Service
IAEP	International Assessment of Educational Progress
I	Sampling Interval
ICR	Inter-Country Coder Reliability Study
ICT	Information Communication Technology
IEA	International Association for the Evaluation of Educational Achievement
INES	OECD Indicators of Education Systems
IRT	Item Response Theory
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISEI	International Socio-Economic Index
MENR	Enrolment for moderately small school
MOS	Measure of size
NCQM	National Centre Quality Monitor
NDP	National Desired Population
NEP	National Enrolled Population
NFI	Normed Fit Index
NIER	National Institute for Educational Research, Japan
NNFI	Non-Normed Fit Index
NPM	National Project Manager
OECD	Organisation for Economic Cooperation and Development
PISA	Programme for International Student Assessment
PPS	Probability Proportional to Size
PGB	PISA Governing Board
PQM	PISA Quality Monitor
PSU	Primary Sampling Units
QAS	Questionnaire Adaptations Spreadsheet
RMSEA	Root Mean Square Error of Approximation
RN	Random Number
SC	School Co-ordinator
SD	Standard Deviation
SEM	Structural Equation Modelling
SMEG	Subject Matter Expert Group
SPT	Study Programme Table
TA	Test Administrator
TAG	Technical Advisory Group
TCS	Target Cluster Size
TIMSS	Third International Mathematics and Science Study
TIMSS-R	Third International Mathematics and Science Study – Repeat
VENR	Enrolment for very small schools
WLE	Weighted Likelihood Estimates

References

Adams, R.J., Wilson, M.R. and **W. Wang** (1997), "The multidimensional random coefficients multinomial logit model", *Applied Psychological Measurement 21,* pp. 1-24.

Aiken, L. R. (1974), "Two scales of attitudes toward mathematics," *Journal for Research in Mathematics Education 5,* National Council of Teachers of Mathematics, Reston, pp. 67-71.

Andersen, Erling B. (1997), "The Rating Scale Model", in van der Linden, W. J. and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory,* Springer, New York/Berlin/Heidelberg.

Bandura, A. (1986), *Social Foundations of Thought and Action: A Social Cognitive Theory,* Prentice Hall, Englewood Cliffs, N.J.

Baumert, J. and **O. Köller** (1998), "Interest Research in Secondary Level I : An Overview", in L. Hoffmann, A. Krapp, K.A. Renninger & J. Baumert (eds.), *Interest and Learning*, IPN, Kiel.

Beaton, A.E. (1987), *Implementing the New Design: The NAEP 1983-84 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

Bryk, A. S. and **S.W. Raudenbush** (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods,* SAGE Publications, Newbury Park.

Bollen, K.A. and **S.J. Long** (eds.) (1993), *Testing Structural Equation Models*, SAGE publications, Newbury Park.

Branden, N. (1994), *Six Pillars of Self-Esteem*. Bantam, New York.

Brennan, R.L. (1992), *Elements of Generalizability Theory*, American College Testing Program, Iowa City.

Buchmann, C. (2000), *Measuring Family Background in International Studies of Educational Achievement: Conceptual Issues and Methodological Challenges,* paper presented at a symposium convened by the Board on International Comparative Studies in Education of the National Academy of Sciences/National Research Council on 1 November, in Washington, D.C.

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

Cronbach, L.J., G.C. Gleser, H. Nanda and **N. Rajaratnam** (1972), *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*, Wiley and Sons, New York.

Eccles, J.S. (1994), "Understanding Women's Educational and Occupational choice: Applying the Eccles *et al.* Model of Achievement-Related Choices", *Psychology of Women Quarterly 18*, Society for the Psychology of Women, Washington, D.C., pp. 585-609.

Eccles, J.S. and **A. Wigfield** (1995), "In the mind of the achiever: The structure of adolescents' academic achievement-related beliefs and self-perceptions", *Personality and Social Psychology Bulletin 21*, Sage Publications, Thousand Oaks, pp. 215-225.

Ganzeboom, H.B.G., P.M. de Graaf and **D.J. Treiman** (1992), "A standard international socio-economic index of occupational status", *Social Science Research 21*, Elsevier, pp.1-56.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley, New York.

Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Grisay, A. (2003), "Translation procedures in OECD/PISA 2000 international assessment", *Language Testing 20,* Holder Arnold Journals, pp.225-240.

Gustafsson, J.E and **P.A. Stahl** (2000), *STREAMS User's Guide, Version 2.5 for Windows,* MultivariateWare, Mölndal, Sweden.

Hacket, G. and **N. Betz.** (1989), "An Exploration of the mathematics Efficacy/mathematics Performance Correspondence", *Journal of Research in Mathematics Education 20,* National Council of Teachers of Mathematics, Reston, pp. 261-273.

Harvey-Beavis, A. (2002), "Student and Questionnaire Development" in OECD, *PISA 2000 Technical Report*, OECD, Paris.

Hatcher, L. (1994), *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling,* SAS Institute Inc., Cary.

International Labour Organisation (1990), *International Standard Classification of Occupations: ISCO-88*, International Labour Office, Geneva.

Jöreskog, K.G. and **Dag Sörbom** (1993), *LISREL 8 User's Reference Guide,* Scientific Software International, Chicago.

Judkins, D.R. (1990), "Fay's Method for Variance Estimation", *Journal of Official Statistics 6,* Statistics Sweden, Stockholm, pp. 223-239.

Kaplan, D. (2000), *Structural Equation Modeling: Foundation and Extensions,* SAGE Publications, Thousand Oaks.

Keyfitz, N. (1951), "Sampling with probabilities proportionate to science: Adjustment for changes in probabilities", *Journal of the American Statistical Association 46,* American Statistical Association, Alexandria, pp.105-109.

Lepper, M. R. (1988), "Motivational considerations in the study of instruction", *Cognition and Instruction 5,* Lawrence Erlbaum Associates, Mahwah, pp. 289-309.

Ma, X. (1999), "A Meta-Analysis of the Relationship Between Anxiety Toward mathematics and Achievement in mathematics", *Journal for Research in Mathematics Education 30,* National Council of Teachers of Mathematics, Reston, pp. 520-540.

Macaskill, G., R.J. Adams and **M.L. Wu** (1998), "Scaling methodology and procedures for the mathematics and science literacy, advanced mathematics and physics scales", in M. Martin and D.L. Kelly (eds.) *Third International Mathematics and Science Study, Technical Report Volume 3: Implementation and Analysis,* Center for the Study of Testing, Evaluation and Educational Policy, Boston College, Chestnut Hill.

Marsh, H. W. (1990), *Self-Description Questionnaire (SDQ) II: A theoretical and Empirical Basis for the Measurement Of Multiple Dimensions of Adolescent Self-Concept: An Interim Test Manual and a Research Monograph,* The Psychological Corporation, San Antonio.

Marsh, H. W. (1994), "Confirmatory factor analysis models of factorial invariance: A multifaceted approach" *Structural Equation Modeling 1,* Lawrence Erlbaum Associates, Mahwah, pp. 5-34.

Marsh, H. W. (1999), *Evaluation of the Big-Two-Factor Theory of Motivation Orientation: Higher-order Factor Models and Age-related Changes*, paper presented at the 31.62 Symposium, Multiple Dimensions of Academic Self-Concept, Frames of Reference, Transitions, and International Perspectives: Studies From the SELF Research Centre. Sydney: University of Western Sydney.

Masters, G. N. and **B. D. Wright** (1997), "The Partial Credit Model", in W. J. van der Linden and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory,* Springer, New York/Berlin/Heidelberg.

Meece, J., A. Wigfield and **J. Eccles** (1990), "Predictors of Maths Anxiety and its Influence on Young Adolescents' Course Enrolment and Performance in Mathematics", *Journal of Educational Psychology 82,* American Psychological Association, Washington, D.C., pp. 60-70.

Middleton, J.A. and **P.A. Spanias** (1999), "Findings, Generalizations, and Criticisms of the Research", *Journal for Research in Mathematics Education 30*, National Council of Teachers of Mathematics, Reston, pp. 65-88.

Mislevy, R.J. (1991), "Randomization-based inference about latent variable from complex samples", *Psychometrika 56*, Psychometric Society, Greensboro, pp. 177-196.

Mislevy, R.J. and **K.M. Sheehan** (1987), "Marginal estimation procedures", in A.E. Beaton (ed.), *The NAEP 1983-1984 Technical Report* (Report No. 15-TR-20), Educational Testing Service, Princeton, N.J.

Mislevy, R.J. and **K.M. Sheehan** (1980), "Information matrices in latent-variable models", *Journal of Educational Statistics 14.4,* American Educational Research Association and American Statistical Association, Washington, D.C., and Alexandria, pp. 335-350.

Mislevy, R.J., A.E. Beaton, B. Kaplan and **K.M. Sheehan.** (1992), "Estimating population characteristics form sparse matrix samples of item responses", *Journal of Educational Measurement 29*, National Council on Measurement in Education, Washington, D.C., pp. 133-161.

Multon, K. D., S. D. Brown and **R.W. Lent** (1991), "Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation", *Journal of Counselling Psychology 38,* American Psychological Association, Washington, D.C., pp. 30-38.

Muthén, B. O., S. H. C. du Toit and **D. Spisic** (1997), "Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical outcomes", *Psychometrika*, Psychometric Society, Greensboro.

Muthen, L. and **B. Muthen** (2003), *Mplus User's Guide Version 3.1,* Muthen & Muthen, Los Angeles.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and its Applications,* University of Toronto Press, Toronto.

OECD (Organisation for Economic Co-Operation and Development) (1999), *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries,* OECD, Paris.

OECD (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, OECD, Paris.

OECD (2002), *PISA 2000 Technical Report,* OECD, Paris.

OECD (2003), *Student Engagement at School: A Sense of Belonging and Participation: Results from PISA 2000,* OECD, Paris.

OECD (2004a), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, OCED, Paris.

OECD (2004b), *Learning for Tomorrow's World – First Results from PISA 2003,* OECD, Paris.

OECD (2004c), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003,* OECD, Paris.

OECD (2005a), *PISA 2003 Data Analysis Manual: SAS® Users,* OECD, Paris.

OECD (2005b), *PISA 2003 Data Analysis Manual: SPSS® Users,* OECD, Paris.

Owens L. and **J. Barnes** (1992), *Learning Preference Scales*, Australian Council for Educational Research, Hawthorn.

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests,* Nielsen and Lydiche, Copenhagen.

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

Rust, K. and **J.N.K. Rao** (1996), "Variance estimation for complex surveys using replication techniques", *Statistical Methods in Medical Research 5,* Holder Arnold Journals, pp. 283-310.

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

Schaffer, E. C., P. S. Nesselrodt and **S. Stringfield** (1994), "The Contribution of Classroom Observation to School Effectiveness Research" in Reynolds *et. al.* (eds.), *Advances in School Effectiveness Research and Practice,* Pergamon, Oxford/New York/Tokyo.

Schulz, W. (2003), *Validating Questionnaire Constructs in International Studies. Two Examples from PISA 2000,* paper presented at the Annual Meeting of the American Educational Research Association (AERA) in Chicago, 21-25 April.

Schulz, W. (2004), "Mapping Student Scores to Item Responses", in W. Schulz and H. Sibberns (eds.), *IEA Civic Education Study. Technical Report*, IEA, Amsterdam.

Sirotnik, K. (1970), "An analysis of variance framework for matrix sampling", *Educational and Psychological Measurement 30,* SAGE Publications, pp. 891-908.

Slavin, R. E. (1983), "When does cooperative learning increase student achievement?" *Psychological Bulletin 94,* American Psychological Association, Washington, D.C., pp. 429-445.

Statistical Solutions (1992), *BMDP Statistical Software*, Statistical Solutions, Los Angeles.

Teddlie, C. and **D. Reynolds** (2000) (eds.), *The International Handbook of School Effectiveness Research,* Falmer Press, London/New York.

Thorndike, R.L. (1973), *Reading Comprehension Education in Fifteen Countries: An Empirical Study*, Almquist & Wiksell, Stockholm.

Travers, K. J., R.A. Garden and **M. Rosier** (1989), "Introduction to the Study", in D.A. Robitaille and R.A. Garden (eds.), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics Curricula,* Pergamon Press, Oxford.

Travers, K. J. and **I. Westbury** (1989), *The IEA Study of Mathematics I: Analysis of Mathematics Curricula,* Pergamon Press, Oxford.

Verhelst, N. (2004), "Generalizability Theory", in Council of Europe, *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, (Section E), Council of Europe (DGIV/EDU/LANG (2004) 13), Strasbourg.

Warm, T. A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika* 54, Psychometric Society, Greensboro, pp. 427-45.

Wigfield, A., J. S. Eccles and **D. Rodriguez** (1998), "The development of children's motivation in school contexts", in P. D. Pearson. and A. Iran-Nejad (eds.), *Review of Research in Education 23,* American Educational Research Association, Washington D.C., pp. 73-118.

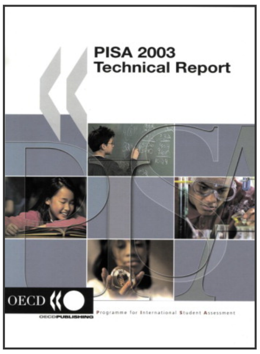
Wilson, M. (1994), "Comparing Attitude Across Different Cultures: Two Quantitative Approaches to Construct Validity", in M. Wilson (ed.), *Objective Measurement II: Theory into Practice,* Ablex, Norwood, pp. 271-292.

Wolter, K.M. (1985), *Introduction to Variance Estimation*, Springer-Verlag, New York.

Wu, M.L., R.J. Adams and **M.R. Wilson** (1997), *ConQuest: Multi-Aspect Test Software* [computer program], Australian Council for Education Research, Camberwell.

Zimmerman, B.J. and **D.H. Schunk** (eds.) (1989), *Self-Regulated Learning and Academic Achievement. Theory, Research and Practice,* Springer, New York.

Table of Contents



From:

PISA 2003 Technical Report

Access the complete publication at:

https://doi.org/10.1787/9789264010543-en

Please cite this chapter as:

OECD (2006), “Survey Weighting and the Calculation of Sampling Variance”, in *PISA 2003 Technical Report*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264010543-9-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.