Australian Council *for* Educational Research

Longitudinal Surveys of Australian Youth

SAMPLING AND WEIGHTING OF THE 2003 LSAY COHORT

Technical Report Number 43

Sheldon Rothman

October 2007

Longitudinal Surveys of Australian Youth: Technical Reports

1. Reading and Numeracy Achievement Tests: 1975-1995, October 1996 (Not publicly available).
2. Overview of the Longitudinal Surveys of Australian Youth Program (LSAY), August 1997
3. The 1961 Cohort Questionnaires: 1975-1994, April 1997
4. The 1965 Cohort Questionnaires: 1981-1995, April 1997
5. The 1970 Cohort Questionnaires: 1985-1994, April 1997
6. The 1975 Cohort Questionnaires: 1989-1996, April 1997
7. The Australian Youth Survey Description
8. Sampling and Samples for the Longitudinal Surveys of Australian Youth
9. Codebook: The LSAY 1995 Year 9 Sample Wave 1 (1995), February 1997
10. Item Statistics for the Reading and Numeracy Tests: 1975-1995
11. Codebook: The LSAY 1995 Year 9 Sample Wave 2 (1996), November, 1997
12. Codebook: the LSAY 1996 School Survey, October 1998
13. Codebook: the LSAY 1996 Teacher Survey, October 1998
14. The Measurement of Socioeconomic Status and Social Class in the LSAY project, November 1999
15. Weighting the 1995 Year 9 Cohort Sample for Differential Response Rates and Sample Attrition, July 2000.
16. The Designed and Achieved Sample of the 1998 LSAY Sample, February 2002
17. Codebook: The LSAY 1995 Year 9 Sample Wave 3 (1997), November 1999
18. Codebook: The LSAY 1995 Year 9 Sample Wave 4 (1998), November 1999
19. Codebook: The LSAY 1998 Year 9 Sample Wave 1 (1998), August 2000
20. Codebook: The LSAY 1999 Teacher Survey for the Y98 cohort, March 2000
21. Codebook: The LSAY 1995 Year 9 Sample Wave5 (1999), June 2000
22. Codebook: The LSAY 1998 Year 9 Sample Wave 2 (1999), September 2000
23. Codebook: The LSAY 1999 School Survey for the Y98 Cohort, November 2000
24. Codebook: The LSAY 1998 Year 9 Sample Wave 3 (2000), April 2001
25. Codebook: The LSAY 1995 Year 9 Sample Wave 6 (2000), May 2001
26. Codebook: The LSAY 1998 Year 9 Sample Wave 4 (2001), June 2002
27. Codebook: The LSAY 1995 Year 9 Sample Wave 7 (2001), March 2003
28. Codebook: The LSAY 1995 Year 9 Sample Wave 8 (2002), May 2003
29. Codebook: The LSAY 1998 Year 9 Sample Wave 5 (2002), May 2003
30. Codebook: The LSAY 1998 Year 9 Sample Wave 6 (2003), May 2004
31. Codebook: The LSAY 1995 Year 9 Sample Wave 9 (2003), May 2004
32. Codebook: The LSAY 1998 Year 9 Sample Wave 7 (2004), May 2005
33. Codebook: The LSAY 1995 Year 9 Sample Wave 10 (2004), May 2005
34. The LSAY 2003 Sample of 15 Year-olds, Wave 1 (2003), December 2005
    Part A: Data Dictionary Part B: Frequencies Part C: Questionnaire
35. The LSAY 2003 Sample of 15 Year-olds, Wave 2 (2004), December 2005
    Part A: Data Dictionary Part B: Frequencies Part C: Questionnaire
36. Codebook: The LSAY 1995 Year 9 Sample Wave 11 (2005), May 2006
37. Codebook: The LSAY 1998 Year 9 Sample Wave 8 (2005), May 2006
38. The LSAY 2003 Sample of 15 Year-olds, Wave 3 (2005), June 2006
    Part A: Data Dictionary Part B: Frequencies Part C: Questionnaire
39. Codebook: The LSAY 1998 Year 9 Sample Wave 9 (2006), April 2007
40. Codebook: The LSAY 1995 Year 9 Sample Wave 12 (2006), April 2007
41. Codebook: The LSAY 2003 Sample of 15 Year-olds, Wave 4 (2006), April 2007
42. Codebook: Preliminary Codebook: The LSAY 2006 Sample of 15 Year Olds Wave 1 (2006), April 2007
43. Sampling and Weighting of the 2003 LSAY Cohort, October 2007

Longitudinal Surveys of Australian Youth

Technical Paper 43

Sampling and Weighting of the 2003 LSAY Cohort

Sheldon Rothman

October 2007

Australian Council *for* Educational Research

ACER

Published 2007 by
Australian Council for Educational Research Ltd
19 Prospect Hill Road, Camberwell, Victoria, 3124, Australia.

Table of Contents

The Designed Sample for PISA 1

Procedures for Information Transfer from PISA to LSAY 1

Calculating the Sample weights 2

Survey Weighting in PISA 2

Survey Weighting 2

Calculating Sampling Variance 6

Adjustments for LSAY 10

References 14

List of Tables

Table 1 Mean scores in the mathematical literacy domain on PISA for five plausible values with different weights 11

Table 2 Mean scores in the reading literacy domain on PISA for five plausible values with different weights 11

Table 3 Mean scores in the scientific literacy domain on PISA for five plausible values with different weights 12

Table 4 Mean scores in the problem solving domain on PISA for five plausible values with different weights 12

Table 5 Distribution on selected variables of full Australian PISA sample with PISA weights and final LSAY sample with LSAY weights 13

this page intentionally left blank

SAMPLING AND WEIGHTING OF THE 2003 LSAY COHORT

The 2003 cohort of the Longitudinal Surveys of Australian Youth (LSAY) was drawn from the sample of 15 year-olds in Australian schools who participated in the Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development (OECD) in 2003. Australia was one of 41 countries that participated in PISA in 2003.

THE DESIGNED SAMPLE FOR PISA

The designed Australian sample for PISA comprised 355 schools from all States and Territories. This sample was designed to be representative of students across Australia, using State/Territory, school sector and region (metropolitan or non-metropolitan) as strata. Within each school, fifty 15 year-olds were selected at random, based on information provided by the school earlier in the year. In schools with fewer than 50 students in the eligible age group, all 15 year-olds were selected. To allow for PISA results to be reported for each State and Territory, the smaller jurisdictions were oversampled. In addition, there was oversampling of Indigenous students so that PISA results could be reliably reported for this group.

Full details of the sampling procedures for PISA can be found in the PISA 2003 Technical Report (OECD, 2005), with specific information for Australia available in the Australian country report (Thomson, Cresswell & De Bortoli, 2004).

There were 12 551 PISA-eligible participants in 321 schools in Australia. This represents a weighted response rate of 90.4 per cent of schools and 83.3 per cent of students. Both rates exceeded the international minimum response rates required for PISA.

PROCEDURES FOR INFORMATION TRANSFER FROM PISA TO LSAY

Students were asked to provide contact details as part of the PISA Student Questionnaire. Contact details for students were transferred from PISA to LSAY, including some students who were later declared ineligible for participation in PISA. Some of the PISA participants had not provided the requested information, some details were not complete and some details were not available because of late returns after follow-up testing in some schools. Letters were sent to the parents of 11 619 who did provide details, advising them of their opportunity to participate in LSAY and the opportunity to withdraw from the survey.

Between August and December 2003, 10 448 young people were successfully contacted and interviewed by the Wallis Consulting Group, the data collection agency contracted for LSAY. The 1171 cases that did not result in successful interviews (10.1% of the cases available) comprised those who provided inaccurate contact details (5.6%), those with whom contact was not made (2.1%) and those who declined to participate (2.4%).

Data matching after the PISA data were processed revealed that 78 of the young people interviewed had not been eligible to participate in PISA, so data from PISA were not available. The main reason for PISA-ineligibility was age. Some schools agreed to participate only if all students in a particular year level were assessed, so young people born outside the PISA-eligible dates of 1 May 1988–30 April 1989 were eliminated from the PISA database. For a small number of cases, the young people contacted for LSAY interviews had not participated in PISA. At the time of the interview, the interviewers were unaware of their non-participation in PISA. These 78 cases have become part of the sub-sample used for pre-testing the questionnaire each year and have been removed from the data files for analysis in LSAY. The final number of eligible participants, who became the LSAY 2003 cohort of 15 year-olds, is 10 370.

CALCULATING THE SAMPLE WEIGHTS

SURVEY WEIGHTING IN PISA[1]

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, 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

[1] Source: Organisation for Economic Co-operation and Development (2005). *PISA 2003 Technical Report*. Paris: OECD. Chapter 8.

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, *Wij*, 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:

$$W_{ij} = t_{2ij} f_{1i} f_{1ij}^{A} t_{1i} w_{2ij} w_{1i}$$

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:

$$w_{1i} = \begin{cases} \dfrac{int(g/i)}{mos(i)} & \text{if } mos(i) < int(g/i) \\ 1 & \text{otherwise} \end{cases}$$

The term *mos (i)* denotes the measure of size given to each school on the sampling frame.

Despite country variations, *mos* (*i*) 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 (i)* = *TCS*.

The term *int (g/i)* denotes the sampling interval used within the explicit sampling stratum *g* that contains school *i* and is calculated as the total of *mos* (*i*) 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 (i)* = 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 int(1/i) = 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 1000 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_{li}, 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.

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-ear-old enrolment exceeded *3 x 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 x 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:

$$w_{2ij} = {enr\,(i)} / {sam\,(i)}$$

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:

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

where the sum in the denominator is over Γ(*i*), the schools within the group (originals and replacements) that participated, while the sum in the numerator is over Ω(*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.

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:

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

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

This procedure 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:

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

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.

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. In Australia, no student's was weight was trimmed at this point in the process (*i.e.* $t_{2ij} < 1.0000$) and one school's base weight was trimmed (*i.e.* $t_{1i} < 1.0000$).

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 tistic 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:

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

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.

ADJUSTMENTS FOR LSAY

Between the time PISA data were gathered in schools and the time LSAY interviews were completed, the unweighted sample size was reduced from 12 551 to 10 370, a loss of 17.4 per cent. This loss of participants was not completely random, necessitating a recalculation of the survey weights as described above. This work was carried out by members of the PISA project team at the Australian Council for Educational Research (ACER) and confirmed by staff at Westat, a partner in the PISA consortium responsible for sample design and weighting.

Weights in PISA are calculated so that the sample of 15 year-olds represents that population, a total of 235 591 for Australia. In contrast, weights have been calculated in previous LSAY cohorts so that the total number of sample members in each year is equal to the number of respondents in that year. To ensure that the 2003 LSAY cohort is consistent with previous LSAY cohorts, the recalculated PISA–LSAY weights were adjusted so that the sum of the 2003 weights is 10 370. This was accomplished by dividing each recalculated PISA weight by 22.7185, or 235 591 divided by 10 370.

Constructing the weights

In the first instance, new sampling weights were constructed to ensure that the distribution of the LSAY sample matched the original PISA sample design across jurisdictions. These new weights ensured the weighted sample sizes were equal (235 591). They also reduced the differences between the full Australian PISA mean scores on the mathematics assessment and the LSAY sample's mean scores on mathematics. An adjustment factor was then constructed, based on the differential attrition that occurred because of the two-month difference between the PISA assessments and the LSAY interviews. This LSAY adjustment factor was based on nine PISA variables that were found to be associated with sample attrition during this period: family structure (FAMSTRUC), the higher level of parents' education (HISCED), country of birth (IMMIG), year level (GRADE), intended occupational level (SSECATEG), education program orientation (ISCEDO), Indigenous background (INDIG), sex (ST03Q01) and home location (LOC).[2] Categories within each of these variables were assigned adjustment factors, with an additional adjustment factor to account for some missing data.

The final LSAY adjustment factor was obtained by finding the product of each of the nine factors and the additional adjustment factor. The LSAY sample distribution weights were then multiplied by the LSAY adjustment factor to obtain the penultimate LSAY 2003 weight. Finally, these weights were reduced so that the total sample size was 10 370, not 235 591.

As a result of these adjustments, the final LSAY sample had mean scores in the four domains higher than the mean scores reported for Australia. Mean scores on the five plausible values in mathematical literacy are shown in Panel A of Table 1 for the full PISA sample and in panel B for the final LSAY sample using the LSAY weights. Table 2 includes the same information for reading literacy, Table 3 for scientific literacy and Table 4 for problem solving.

Distributions of the 2003 LSAY cohort on selected variables are presented in Table 5, with comparisons to the 2003 Australian PISA sample. Retention across most groups is reasonable, although there is some evidence of nonrandom attrition that is not fully

[2] The PISA variables (also used in LSAY) are designated by uppercase letters in parentheses. In the LSAY data set sex is represented by the variable SEX, which corrects six cases in the original PISA data.

ameliorated by the weights that have been applied. Annual weights should ensure that bias caused by nonrandom attrition is minimised. It may also be necessary to recalculate weights for 2003, if nonrandom attrition in subsequent waves of the data collection shows other factors not included in the weighting procedures described in this technical paper.

Table 1 Mean scores in the mathematical literacy domain on PISA for five plausible values with different weights

Panel A Full Australian PISA sample with PISA weights

	pv1math Plausible value in math	pv2math Plausible value in math	pv3math Plausible value in math	pv4math Plausible value in math	pv5math Plausible value in math
Mean	524.084088	524.194639	524.595756	524.461881	523.993657
N	235591	235591	235591	235591	235591
Std. Deviation	95.5986418	95.2732767	95.1469202	95.5083092	95.5808836

Panel B Final LSAY sample with LSAY weights

	pv1math Plausible value in math	pv2math Plausible value in math	pv3math Plausible value in math	pv4math Plausible value in math	pv5math Plausible value in math
Mean	529.120088	529.305477	529.675114	529.103138	528.529694
N	10370	10370	10370	10370	10370
Std. Deviation	93.7286359	93.1343523	93.2072986	93.6633141	93.9642639

Table 2 Mean scores in the reading literacy domain on PISA for five plausible values with different weights

Panel A Full Australian PISA sample with PISA weights

	pv1read Plausible value in reading	pv2read Plausible value in reading	pv3read Plausible value in reading	pv4read Plausible value in reading	pv5read Plausible value in reading
Mean	525.667475	525.042760	525.173378	525.936593	525.314827
N	235591	235591	235591	235591	235591
Std. Deviation	97.1966627	96.7954536	97.9065269	97.7490983	97.2343510

Panel B Final LSAY sample with LSAY weights

	pv1read Plausible value in reading	pv2read Plausible value in reading	pv3read Plausible value in reading	pv4read Plausible value in reading	pv5read Plausible value in reading
Mean	530.011458	529.692670	529.950200	530.407197	530.105002
N	10370	10370	10370	10370	10370
Std. Deviation	94.4527844	93.9720446	94.6964048	95.1272016	94.3057013

Table 3 Mean scores in the scientific literacy domain on PISA for five plausible values with different weights

Panel A Full Australian PISA sample with PISA weights

	pv1scie Plausible value in science	pv2scie Plausible value in science	pv3scie Plausible value in science	pv4scie Plausible value in science	pv5scie Plausible value in science
Mean	525.378123	524.728628	524.856084	525.380993	524.928667
N	235591	235591	235591	235591	235591
Std. Deviation	101.7393159	101.6761871	101.8559128	101.8864632	101.9976304

Panel B Final LSAY sample with LSAY weights

	pv1scie Plausible value in science	pv2scie Plausible value in science	pv3scie Plausible value in science	pv4scie Plausible value in science	pv5scie Plausible value in science
Mean	530.245586	529.968106	530.467392	530.404869	530.031822
N	10370	10370	10370	10370	10370
Std. Deviation	100.1752441	99.4773224	99.4873180	99.6501463	100.3380113

Table 4 Mean scores in the problem solving domain on PISA for five plausible values with different weights

Panel A Full Australian PISA sample with PISA weights

	pv1prob Plausible value in problem solving	pv2prob Plausible value in problem solving	pv3prob Plausible value in problem solving	pv4prob Plausible value in problem solving	pv5prob Plausible value in problem solving
Mean	529.905483	529.489230	530.195116	529.552225	530.086995
N	235591	235591	235591	235591	235591
Std. Deviation	91.2870140	91.3364657	90.9430723	91.5325777	91.7632754

Panel B Final LSAY sample with LSAY weights

	pv1prob Plausible value in problem solving	pv2prob Plausible value in problem solving	pv3prob Plausible value in problem solving	pv4prob Plausible value in problem solving	pv5prob Plausible value in problem solving
Mean	535.132395	534.455995	535.451331	534.653024	535.128858
N	10370	10370	10370	10370	10370
Std. Deviation	89.3064129	89.4502706	88.8235453	89.0427699	90.0426785

Table 5 Distribution on selected variables of full Australian PISA sample with PISA weights and final LSAY sample with LSAY weights

		PISA		LSAY	
Variable	**Variable categories**	**Weighted N**	**Per cent**	**Weighted N**	**Per cent**
Total		235591	100.0	10370	100.0
Sex	1 Female	115829	49.2	5097	49.2
	2 Male	119762	50.8	5273	50.8
State/Territory	1 ACT	4449	1.9	196	1.9
	2 NSW	74568	31.7	3282	31.7
	3 VIC	56849	24.1	2502	24.1
	4 QLD	45385	19.3	1998	19.3
	5 SA	21089	9.0	928	9.0
	6 WA	26193	11.1	1153	11.1
	7 TAS	5292	2.2	233	2.2
	8 NT	1766	0.7	78	0.7
MCEETYA Location Class	1 Metropolitan Zone Mainland State Capital City regions	142696	60.6	6196	59.7
	2 Metropolitan Zone Major urban Statistical Districts	27525	11.7	1243	12.0
	3 Provincial Zone Inner provincial areas	27674	11.7	1232	11.9
	4 Provincial Zone Outer provincial areas	13353	5.7	613	5.9
	5 Provincial Zone Provincial City Statistical Districts 25,000	10228	4.3	464	4.5
	6 Provincial Zone Provincial City Statistical Districts 50,000	12661	5.4	558	5.4
	7 Remote Zone Remote areas	1340	0.6	59	0.6
	8 Remote Zone Very Remote areas	114	0.0	5	0.0
Indigenous Status	0 Non-Indigenous	230398	97.8	10175	98.1
	1 Indigenous	5193	2.2	195	1.9
Family Structure*	1 Single parent family	46649	20.0	1924	18.7
	2 Nuclear family	161819	69.4	7341	71.2
	3 Mixed family	18342	7.9	817	7.9
	4 Other	6355	2.7	223	2.2
Highest parent White collar/ Blue collar classification*	1 White collar high skilled	143538	65.4	6399	65.6
	2 White collar low skilled	40429	18.4	1809	18.5
	3 Blue collar high skilled	18968	8.6	835	8.6
	4 Blue collar low skilled	16534	7.5	719	7.4
Highest educational level of parents*	0 None	3244	1.4	143	1.4
	1 ISCED 1	1309	0.6	44	0.4
	2 ISCED 2	25843	11.3	1115	11.0
	3 ISCED 3B, C	5222	2.3	232	2.3
	4 ISCED 3A, ISCED 4	69076	30.2	3058	30.2
	5 ISCED 5B	32665	14.3	1486	14.7
	6 ISCED 5A, 6	91341	39.9	4033	39.9
Country of birth*	1 Native students	177966	77.3	7949	77.9
	2 First-Generation students	27034	11.7	1208	11.8
	3 Non-native students	25318	11.0	1047	10.3
Grade	-3 Year 7	15	0.0	1	0.0
	-2 Year 8	335	0.1	12	0.1
	-1 Year 9	19639	8.3	875	8.4
	0 Year 10	170233	72.3	7494	72.3
	+1 Year 11	45247	19.2	1982	19.1
	+2 Year 12	123	0.1	6	0.1
Expected educational level of student (ISCED)*	0 None	928	0.4	29	0.3
	1 ISCED 2	6243	2.7	241	2.3
	2 ISCED 3B, C	8614	3.7	355	3.4
	3 ISCED 3A, ISCED 4	53297	22.7	2345	22.6
	4 ISCED 5B	18734	8.0	824	8.0
	5 ISCED 5A, 6	146576	62.5	6566	63.4
Self White collar/Blue collar classification*	1 White collar high skilled	144797	76.0	6518	76.4
	2 White collar low skilled	23798	12.5	1042	12.2
	3 Blue collar high skilled	20254	10.6	902	10.6
	4 Blue collar low skilled	1688	0.9	71	0.8

Note: Variables marked with * do not show missing values. Counts do not sum to totals, but percentages sum to 100%. All values are rounded, so some variables may not sum to 100.0%.

REFERENCES

Organisation for Economic Co-operation and Development (2005). *PISA 2003 Technical Report*. Paris: OECD.

Thomson, S., Cresswell, J. & De Bortoli, L. (2004). *Facing the future: A focus on mathematical literacy among Australian 15-year-old students in PISA 2003*. Melbourne: ACER.