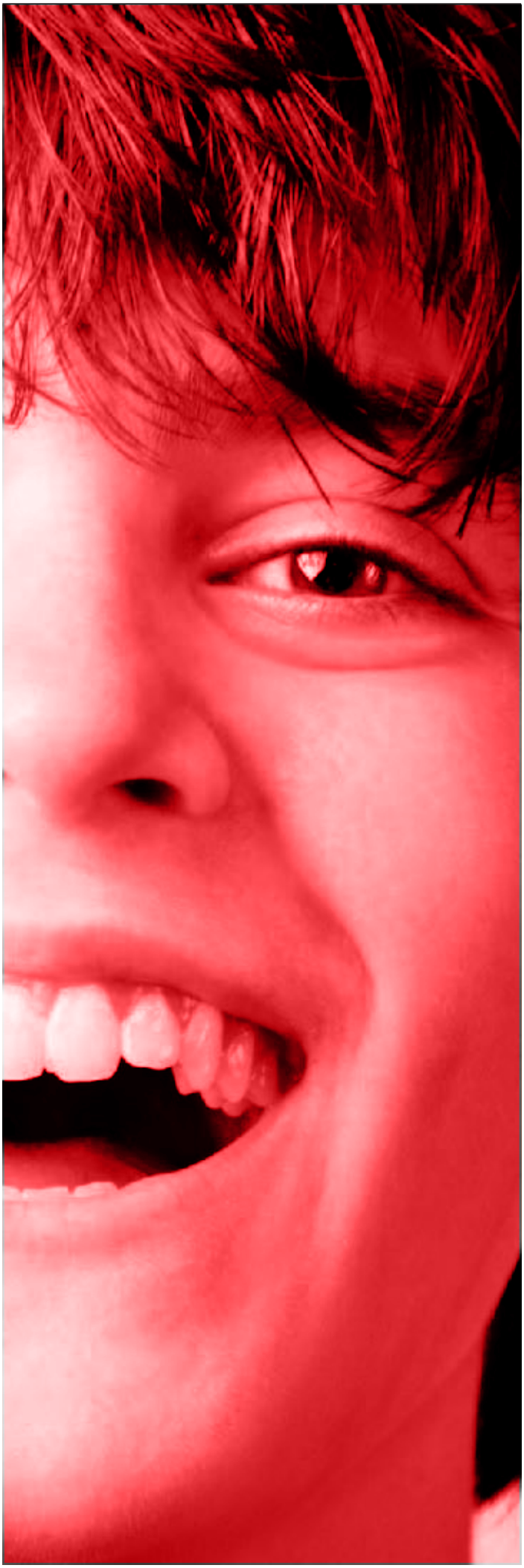

LONGITUDINAL SURVEYS OF AUSTRALIAN YOUTH

TECHNICAL REPORT 61

# Weighting the LSAY Programme of International Student Assessment cohorts

Patrick Lim

National Centre for Vocational Education Research

Longitudinal Surveys of Australian Youth

NCVER

# Weighting the LSAY Programme of International Student Assessment cohorts

Patrick Lim
NCVER

NATIONAL CENTRE FOR VOCATIONAL EDUCATION RESEARCH

**TECHNICAL REPORT 61**

This document should be attributed as Lim 2011, *Weighting the LSAY Programme of International Student Assessment cohorts,* NCVER.

This work has been produced by NCVER through the Longitudinal Surveys of Australian Youth (LSAY) Program, on behalf of the Australian Government and state and territory governments, with funding provided through the Australian Department of Education, Employment and Workplace Relations.



ISBN 978 1 921955 75 4 web edition
978 1 921955 76 1 print edition
TD/TNC 105.07

Published by NCVER
ABN 87 007 967 311

Level 11, 33 King William Street, Adelaide SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

**P** +61 8 8230 8400 **F** +61 8 8212 3436 **E** ncver@ncver.edu.au **W** <http://www.ncver.edu.au>

# About the research

## *Weighting the LSAY Programme of International Student Assessment cohorts*

Patrick Lim, NCVER

The 2003 and 2006 cohorts of the Longitudinal Surveys of Australian Youth (LSAY) are derived from the 2003 and 2006 Programme of International Student Assessment (PISA). LSAY continues to survey these individuals for approximately ten years after their participation in PISA.

The Programme of International Student Assessment uses a stratified sample scheme to sample individuals, with sample weights created in PISA to ensure that the resultant sample represents the underlying population of interest. The longitudinal nature of LSAY means that over time individuals drop out of the sample. The original sample weights must therefore be adjusted to account for differential attrition to ensure that the LSAY sample in each wave continues to represent the underlying population.

This technical note outlines the methodology used to adjust the original weights to ensure that this occurs. It also provides guidance to researchers for applying the weights to their analysis of LSAY data.

Tom Karmel
Managing Director, NCVER

# Contents


Tables 6

Introduction 7

Methodology 8

- PISA sampling scheme 8
- Stratification 9
- From PISA to LSAY 10
- LSAY weights 11
- Distribution of weights — Y03 and Y06 17

Recommendations for applying weights 19

References 21

Appendices

- A: Sample weights 22
- B: Attrition weights 24

# Tables


| | | |
|---|---|---|
| 1 | PISA population and sample information | 9 |
| 2 | Number of schools selected by strata, PISA 2003 | 10 |
| 3 | Number of schools selected by strata, PISA 2006 | 10 |
| 4 | Weighting variables, Y03 and Y06 | 12 |
| 5 | Y03 weights, 2003 and 2009 | 14 |
| 6 | Y03 weights — achievement scores, 2003 and 2009 | 15 |
| 7 | Y06 weights, 2006 and 2009 | 16 |
| 8 | Summary statistics for Y03 weights | 17 |
| 9 | Summary statistics for Y06 weights | 18 |
| A1 | Y03 sample weights, 2003 and 2009 | 22 |
| A2 | Y06 sample weights, 2006 and 2009 | 23 |
| B1 | Y03 attrition weights, 2003 and 2009 | 25 |
| B2 | Y03 attrition weights — achievement scores, 2003 and 2009 | 26 |
| B3 | Y06 attrition weights, 2006 and 2009 | 27 |

# Introduction

This technical note outlines the methodology used for creating the weights in the 2003 and 2006 cohorts of the Longitudinal Surveys of Australian Youth (LSAY). This document accompanies the LSAY user guides as a reference for researchers (NCVER 2011a, 2011b).

To rebalance the data, weights are applied to survey data, thereby ensuring that the sample represents the original population. Rebalancing is necessary because of the selection of individuals with unequal probability and unit non-response. Each respondent in the survey receives a weight. Individuals who represent under-represented groups are allocated larger weights, and those who represent over-represented groups are allocated smaller weights.

Longitudinal surveys have an added level of complexity, as they suffer from attrition. The weights created in LSAY attempt to overcome the effects of different rates of non-response over time from different groups.

The weights created in LSAY attempt to ensure that the sample matches the original population, given both the use of a complex sampling scheme and attrition. In most analytical techniques (cross-tabulations, regressions etc.), weights need to be applied to ensure that the results obtained reflect the original population. For more complicated techniques such as longitudinal data analysis (panel analysis), researchers need to consider the impact of the survey design and attrition and determine an appropriate methodology.

This paper begins with a brief discussion of the sampling methodology, followed by descriptions of how the weights are calculated for the 2003 and 2006 cohorts. The final section contains some recommendations on the use of weights when analysing survey data.

# Methodology

This section outlines the sampling and weighting methodology used for the 2003 (Y03) and 2006 (Y06) PISA samples. Full descriptions of the sampling scheme, sampling methodology and weighting can be found in OECD (2005, 2009). Further information regarding details for the Australian sample can be found in Thomson, Creswell and De Bortoli (2004) and Thomson and De Bortoli (2008).

Given that the 2003 and 2006 PISA surveys comprise the first wave of the LSAY surveys, the sampling methodology for LSAY is exactly that used in the Programme of International Student Assessment (PISA). In particular, the Organisation for Economic Co-operation and Development (OECD) has derived sample weights for the PISA waves (wave 1) in both cohorts. The LSAY component of PISA was conducted as a telephone interview carried out several months after the PISA sample; this is the methodology used for the 2003 cohort and this telephone interview sample makes up the first wave of the LSAY 2003 sample. However, this sampling methodology has resulted in attrition from the original PISA sample, which means there is a need to adjust the PISA sampling weight in wave 1 to ensure that this LSAY sample matches the PISA population. In the 2006 cohort, the LSAY questions were asked as part of the 2006 PISA survey, and so the sample weights in wave 1 are those generated by the OECD.

In subsequent waves of LSAY, the derived weights incorporate both the original sampling scheme and weights and also account for the effects of attrition.

## PISA sampling scheme

The PISA target population is 15-year-old students attending educational institutions and in Year 7 or higher (noting that the modal age for both cohorts was Year 10). Part-time students, students undertaking only vocational education and training (VET), and students attending foreign schools are excluded. Further exclusions include those who are schooled at home, in the workplace or out of the country. The international age requirement is that an individual had to be 15 years old during the period of 1 March 2003 to 31 August 2003, or 1 March 2006 to 31 August 2006 for Y03 and Y06 respectively.

PISA is a two-stage stratified sample. The first stage comprises the sampling of individual schools, while in the second stage individual 15-year-old students in each of the designated schools are sampled.

Schools are sampled using probability proportional to enrolment size of 15-year-olds (PPS). The following short example shows how this occurs.

> Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 15-year-old students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on, to the last school (1011 to 1500). We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137; that is, the first, fourth, and sixth schools.

Under this sampling scheme, larger schools have a greater chance of selection than smaller schools.

The international minimum sample sizes required for the Programme of International Student Assessment were 150 schools and 4500 students. In Australia, more schools were sampled to enable comparisons between jurisdictions. The Australian Council for Educational Research (ACER) developed the sampling frame, using Australian Bureau of Statistics (ABS) data and other jurisdictional sources. The sampling frame excluded correctional, offshore and very remote mainland schools, schools teaching in a language other than English, and VET institutions. The choice to exclude VET institutions is due to practical considerations because the total population of 15-year-olds attending is very small.

Once the schools have been selected, the second stage is to select individual 15-year-olds from each of the schools. The PISA sampling guide requires that 35 individuals per school are selected. If a school has fewer than 35, then all 15-year-old students are selected. Where schools have more than 35 students aged 15, individuals are selected with equal probability.

Excluded from selection were students with a severe physical or sensory disability, with a severe intellectual or emotional disability, or with limited proficiency in English (defined as having received less than one year of instruction in English). Less than 1% of selected 15-year-olds were excluded due to these criteria. All Indigenous 15-year-olds at each school were selected to participate in PISA. In 2003 and 2006, 1300 and 1080 Indigenous students respectively participated in PISA.

The total number of students participating in PISA was 12 551 and 14 170 for the 2003 and 2006 cohorts respectively. Table 1 describes the target population and sample for the 2003 and 2006 PISA samples.

**Table 1 PISA population and sample information**

| | Cohort | |
|---|---|---|
| | 2003 | 2006 |
| Total population of 15-year-olds | 268 164 | 270 115 |
| Total enrolled population of 15-year-olds | 250 635 | 256 754 |
| Total in national desired target population | 248 035 | 255 554 |
| School-level exclusions | 1 615 | 1 371 |
| Total in national desired target population after school exclusions and before within-school exclusions | 246 420 | 254 183 |
| Percentage of school-level exclusions | 0.65 | 0.54 |
| Number of participating students | 12 551 | 14 170 |
| Weighted number of participating students | 235 591 | 234 940 |
| Number of excluded students | 228 | 234 |
| Weighted number of excluded students | 3 612 | 2 935 |
| Within-school exclusion rate (%) | 1.51 | 1.23 |
| Overall exclusion rate (%) | 2.15 | 1.76 |
| Coverage Index 1: coverage of national desired population | 0.98 | 0.98 |
| Coverage Index 2: coverage of national enrolled population | 0.97 | 0.98 |
| Coverage Index 3: coverage of 15-year-old population | 0.88 | 0.87 |

Source: OECD (2005, table A3.1); Thompson & De Bortoli (2008, table A2.3).

## Stratification

The second aspect to the sampling is the use of stratification. In both the 2003 and 2006 Programme of International Student Assessment cohorts, stratification was used to:

- improve the efficiency of the sample design
- make sure that all parts of a population were included in a sample (for example, states, sectors)
- ensure adequate representation of specific groups of the target population in the sample.

The stratification variables used in Australia were state/territory, school sector (independent, Catholic/government), and school location (metropolitan/country). In the sampling process, schools are ordered by their size within their strata. Individual schools are then selected using probability proportional to size within their strata group.

The number of schools selected to participate in the 2003 and 2006 PISA samples is presented in tables 2 and 3 (by state and sector).

**Table 2 Number of schools selected by strata, PISA 2003**

| State/territory | Catholic | Government | Independent | Total |
|---|---|---|---|---|
| NSW | 21 | 58 | 11 | 90 |
| Vic. | 14 | 39 | 11 | 64 |
| QLD | 10 | 35 | 10 | 55 |
| SA | 7 | 20 | 7 | 34 |
| WA | 8 | 27 | 7 | 42 |
| Tas. | 4 | 15 | 2 | 21 |
| NT | 3 | 12 | 4 | 19 |
| ACT | 7 | 20 | 3 | 30 |
| **Total** | **74** | **226** | **55** | **355** |

Source: Thomson, Creswell & De Bortoli (2004).

In the 2003 cohort, there were 355 schools selected to participate in PISA. Of these, 45 chose not to participate. Eleven schools were added as replacement schools; the total number of schools participating in PISA in 2003 was 321. This represents an overall school response rate of 90.4%.

**Table 3 Number of schools selected by strata, PISA 2006**

| State/territory | Catholic | Government | Independent | Total |
|---|---|---|---|---|
| NSW | 8 | 15 | 3 | 26 |
| Vic. | 20 | 50 | 13 | 83 |
| QLD | 12 | 35 | 11 | 58 |
| SA | 11 | 38 | 10 | 59 |
| WA | 8 | 27 | 9 | 44 |
| Tas. | 8 | 24 | 9 | 41 |
| NT | 5 | 25 | 5 | 35 |
| ACT | 4 | 16 | 7 | 27 |
| **Total** | **76** | **230** | **67** | **373** |

Source: Thompson and De Bortoli (2008).

In 2006, there were 373 schools selected to participate. Of these, 16 schools were not eligible and an additional two did not participate. One of these schools was replaced by another school, which meant 356 schools participating in the 2006 PISA. This represents an overall school response rate of 95.4%.

## From PISA to LSAY

As part of the PISA questionnaire, students were asked to provide their contact details. For the 2003 cohort, these contact details were used in a follow-up phone interview for LSAY in 2003. Of the 12 551 participants in 2003, only 10 448 were successfully contacted and interviewed. Of these, 78 were ineligible for PISA (age or other factors) and so the total number of individuals who completed both the PISA and LSAY questionnaires was 10 370. These 10 370 individuals comprise the first wave of

the 2003 LSAY sample. The PISA data for all 12 551 individuals are available and provide valuable information for the creation of weights.

For the 2006 cohort, the LSAY questions were included in the PISA questionnaire, and so all 14 170 individuals who participated in PISA form the LSAY cohort. Again, these individuals were asked their contact details to enable follow-up interviews from 2007 onwards.

The attrition in the first wave of the 2003 cohort means there is a need to adjust the PISA sample weights to account for the attrition that occurred between the PISA and LSAY surveys.

## LSAY weights

A sample survey is designed to represent a population of interest. In LSAY's case, this population is the number of 15-year-olds attending school (or other similar institution) during the period 1 March to 31 August in the relevant PISA survey year. As the Programme of International Student Assessment uses a two-stage stratified sample, individual schools and students are selected with uneven probabilities. In particular, larger schools have a higher chance of selection than smaller schools. Weights are created to ensure that the selected sample(s) match the original population. There may be a higher proportion of schools sampled from New South Wales, for example, than occurs in the original population. In this case, the survey weights would weight down all schools from NSW so that they match the distribution in the original population. Conversely, those schools (or individuals) that are under-represented in the selected sample would be weighted up.

In wave 1 (PISA), the sample weights were derived by the PISA consortium. The weights are based on the sampling scheme employed and the probability of selection of a school and an individual. The weights are constructed to ensure that, when applied, the collected sample represents the underlying population of 15-year-olds attending school. The methodology for developing the sampling weights for PISA appears in each of the PISA technical manuals (OECD 2005, 2009) and is not repeated here.

The National Centre for Vocational Education Research (NCVER) does not have access to the full population information for the PISA sample when creating the weights for LSAY, so it is assumed that the population totals and distributions are those obtained by applying the PISA weights to the full PISA datasets.

The LSAY sample suffers from year-on-year attrition. In the case of Y03, this attrition begins in the first wave and continues for each year of surveying. Year-on-year attrition is also observed for the Y06 cohort. Attrition means that the PISA sample weights are no longer representative of the population and they therefore need to be recalculated. In addition, because different groups of people tend to drop out of the survey at differing rates, the weights are further adjusted to ensure that each wave of the LSAY sample matches the original PISA population in relation to a given set of background and sampling variables.

That is, the final weights incorporate adjustments for both the sampling scheme employed and the effects of attrition.

The LSAY weights are created by adjusting the PISA sampling weights by the inverse probability of responding in a given wave. There are several ways of determining this probability, such as simple cross-tabulations or proportions. The approach that we have used however is logistic regression, an

approach that allows us to determine the probability of response for a large number of explanatory variables. The response variable of interest is a binary variable, such that:

$$Y_i = \begin{cases} 1, if\ an\ individual\ responded\ in\ the\ given\ wave \\ 0, if\ an\ individual\ did\ not\ respond\ in\ the\ given\ wave \end{cases}$$

The explanatory variables used in the regression include those used in selecting the sample (and defining the strata), state and school sector, as well as those that contribute to differential attrition. Table 4 shows the variables used for determining the probability of responding for both cohorts.

**Table 4 Weighting variables, Y03 and Y06**

| Y03* | | Y06 | |
|---|---|---|---|
| PISA variables** | Description | PISA variables** | Description |
| STATE | State of school attending | STATE | State of school attending |
| SECTOR | Sector of school | SECTOR | Sector of school |
| FAMSTRUC | Family structure | INDIG* | Indigenous status |
| HISCED | Highest parental education | HISCED | Highest educational level of parents |
| GRADE*** | Student year level, relative to modal school year | GRADE*** (ST01Q01) | Student year level, relative to modal school year |
| GENDER (ST03Q01) | Gender | GENDER (ST04Q01) | Gender |
| Immigration status (IMMIG) | Immigration status | Immigration Status (AUSIMMIG) | Immigration status |
| SSECATEG | Occupational status of parents | GEOLOC_3 | Geographic location of school attended in 2006 |
| Mathematics achievement (pv1math – pv5math) | Each of the 5 plausible values included in the regression | Mathematics achievement (pv1math_q) | Mathematics achievement quartile used on the first plausible value included in regression |
| Reading achievement (pv1read – pv5read) | Each of the 5 plausible values included in the regression | | |
| Science achievement (pv1scie – pv5scie) | Each of the 5 plausible values included in the regression | | |
| Problem-solving achievement (pv1prob – pv5prob) | Each of the 5 plausible values included in the regression | | |

Notes: * In deriving attrition weights, it is important that the included variables are those that were asked as part of PISA. In 2003, Indigenous status was asked only of LSAY respondents. Thus, it is not possible to include this as an attrition weighting variable.

** The weights for each of the cohorts were created at different times. Thus, the included variables are different across the cohorts. It is anticipated that a future revision of the weights will address this issue, in particular, the inclusion of further achievement variables in Y06.

*** PISA samples 15-year-olds regardless of their school year level; this variable indicates an individual's school year level relative to Year 10.

Two weights are created for each wave of the LSAY data; they are weights that return:

1 the sample size for the given wave (n) (normalised weight) [1]

2 the PISA population total (N).

The use of either weight will ensure that the distributions of the variables used in creating the weights will be similar to the distributions of these variables in the original weighted PISA population.

[1] The normalised weight is useful for those occasions when researchers have no access to software that correctly implements survey weighting methodology. This weight ensures that the standard errors and inference calculations are undertaken using the number of observations in the sample rather than the weighted N. Further, in earlier LSAY cohorts, weights have been constructed so that the total number of sample members in each wave is equal to the number of respondents in that year. To ensure that the 2003 LSAY cohort is consistent with previous LSAY cohorts, the weights are adjusted so that the sum of the weights is equal to the sample size in the respective wave (Rothman 2007).

The normalised weights are useful when researchers use statistical software that associates the sum of the weights with the number of observations. This results in an appearance of much more statistical power than is actually present.

## Logistic regression

The logistic model used to generate the probability of responding in each wave is

$$logit(\boldsymbol{y}) = \boldsymbol{\alpha} + \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is an indicator variable[2], with the figures 1 and 0 representing response and non-response for a given survey wave, $\boldsymbol{\beta}$ is the vector of regression coefficients as given in table 4, $\boldsymbol{X}$ is the design matrix for relevant covariates and $\boldsymbol{\varepsilon}$ represents the random error component.

The probability of an individual responding in a given wave is:

$$\hat{p}_i = \frac{e^{\hat{\alpha}+\hat{\beta}X}}{1 + e^{\hat{\alpha}+\hat{\beta}X}}$$

Given this probability, an interim weight for each individual is derived using the inverse probability, $\frac{1}{\hat{p}_i}$. In order to construct the weights provided in LSAY, there are a further two adjustments made to this interim weight.

1 The inverse probability is multiplied by the original PISA sampling weight, such that,

$$wt_i = PISA\_wt_i \times (\frac{1}{\hat{p}_i}),$$

for all individuals who responded in the given survey wave.

2 The weight ($wt_i$) is then multiplied by one of the two following constants to create the population and normalised weights:

a. Population weight,

$$adj_N = \frac{\sum_{i=1}^{N} PISA\_wt_i}{\sum_{i=1}^{n} wt_i},$$

where N represents the original PISA sample size, and n is the sample size in the given LSAY wave.

b. Normalised weight,

$$adj_n = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} wt_i},$$

where n is the sample size in the given LSAY wave.

The two weights created for each LSAY wave are finally derived using:

$$wt_{iN} = wt_i \times adj_N$$
$$wt_{in} = wt_i \times adj_n$$

A listing of the weighting variables available in the LSAY datasets is available in the LSAY user guides (NCVER 2011a, 2011b).[3] Historically, the LSAY weight variables have been constructed by first recalculating the sampling weights and the attrition weights separately. These two weights are then multiplied. The wave-on-wave sample and attrition weights have been derived using the logistic

[2] The indicator variable is present for each survey wave. The variable in the datasets is inYYYY, where YYYY is the wave of interest.

[3] In particular, the weighting variables are labelled as wtYYYY for normalised weights, and wtYYYY_P for population weights. PISA weighting variables are w_fstuwt in both cohorts.

regression approach presented above and are available in the LSAY datasets. Further details on these weights are available in appendices A and B.

The following section details the impact of using the LSAY weights for the latest wave of data (the 2009 wave). The 2009 wave has been selected as it is the wave that presently suffers from the greatest extent of attrition. The tables presented below use the population weights. The results for the normalised weights are identical, with the exception that the sample size is returned as the total rather than the population total. The results for other waves are similar (not shown).

## Y03 weights

Tables 5 and 6 present the effects of the weights on the Y03 data in the 2003 and 2009 waves of data for the variables used to create the weights.

**Table 5 Y03 weights, 2003 and 2009**

| | PISA 2003 | | LSAY Y03 2003 | | LSAY Y03 2009** | |
|---|---|---|---|---|---|---|
| *Variable* | Unweighted % | Population* % | Unweighted % | Weighted % | Unweighted % | Weighted % |
| **State** | | | | | | |
| ACT | 7.12 | 1.89 | 7.00 | 1.89 | 7.82 | 1.86 |
| NSW | 23.78 | 31.65 | 22.79 | 31.75 | 22.89 | 32.00 |
| Vic. | 18.76 | 24.13 | 19.48 | 24.14 | 20.02 | 24.38 |
| QLD | 15.41 | 19.26 | 15.73 | 19.05 | 15.49 | 19.02 |
| SA | 9.83 | 8.95 | 10.02 | 8.90 | 10.06 | 8.61 |
| WA | 14.08 | 11.12 | 14.29 | 11.18 | 13.44 | 11.22 |
| Tas. | 6.41 | 2.25 | 6.56 | 2.24 | 6.65 | 2.21 |
| NT | 4.65 | 0.75 | 4.14 | 0.75 | 3.63 | 0.71 |
| **Sector** | | | | | | |
| Government | 64.58 | 61.84 | 64.06 | 61.72 | 60.00 | 61.83 |
| Catholic | 19.62 | 21.14 | 20.59 | 21.15 | 21.79 | 21.15 |
| Independent | 15.79 | 17.03 | 15.35 | 17.13 | 18.21 | 17.02 |
| **Family structure (FAMSTRUC)** | | | | | | |
| Single parent family | 21.17 | 19.80 | 19.97 | 19.85 | 16.24 | 20.21 |
| Nuclear family | 66.72 | 68.69 | 68.94 | 68.83 | 74.41 | 68.53 |
| Mixed family | 8.10 | 7.79 | 8.00 | 7.78 | 6.74 | 7.75 |
| Other | 2.88 | 2.70 | 2.47 | 2.63 | 2.03 | 2.63 |
| Missing/unknown | 1.13 | 1.03 | 0.62 | 0.92 | 0.58 | 0.88 |
| **Highest educational level of parents (HISCED)** | | | | | | |
| None | 1.25 | 1.38 | 1.21 | 1.39 | 1.02 | 1.35 |
| ISCED 1 | 0.5020 | 0.56 | 0.39 | 0.52 | 0.29 | 0.57 |
| ISCED 2 | 11.19 | 10.97 | 10.98 | 11.00 | 8.91 | 11.22 |
| ISCED 3B, C | 2.36 | 2.22 | 2.36 | 2.23 | 1.55 | 2.29 |
| ISCED 3A, 4 | 29.04 | 29.32 | 29.46 | 29.29 | 27.18 | 29.53 |
| ISCED 5B | 13.42 | 13.87 | 13.55 | 13.89 | 13.11 | 13.65 |
| ISCED 5A, 6 | 39.21 | 38.77 | 39.75 | 38.85 | 46.43 | 38.52 |
| Missing/unknown | 3.04 | 2.93 | 2.30 | 2.85 | 1.60 | 2.87 |

| | PISA 2003 | | LSAY Y03 2003 | | LSAY Y03 2009** | |
|---|---|---|---|---|---|---|
| *Variable* | Unweighted % | Population* % | Unweighted % | Weighted % | Unweighted % | Weighted % |
| **Grade (relative to modal grade, Year 10)** | | | | | | |
| -3 | 0.008 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |
| -2 | 0.13 | 0.14 | 0.11 | 0.15 | 0.05 | 0.24 |
| -1 | 8.48 | 8.34 | 8.37 | 8.31 | 8.33 | 8.52 |
| 0 | 71.56 | 72.26 | 71.15 | 72.31 | 71.14 | 71.80 |
| 1 | 19.76 | 19.21 | 20.30 | 19.17 | 20.38 | 19.38 |
| 2 | 0.06 | 0.05 | 0.07 | 0.05 | 0.09 | 0.05 |
| **Sex (ST03Q01)** | | | | | | |
| Male | 49.53 | 49.17 | 50.39 | 49.18 | 49.97 | 49.68 |
| Female | 50.47 | 50.83 | 49.61 | 50.82 | 50.03 | 50.32 |
| **Immigration status (IMMIG)** | | | | | | |
| Native students | 77.14 | 75.54 | 78.42 | 75.58 | 78.10 | 75.23 |
| First generation students | 10.69 | 11.48 | 10.65 | 11.59 | 10.76 | 11.53 |
| Non-native students | 10.02 | 10.75 | 9.40 | 10.75 | 9.88 | 11.13 |
| Missing/unknown | 2.14 | 2.24 | 1.53 | 2.08 | 1.26 | 2.11 |
| **Occupational status of parents (SSECATEG)** | | | | | | |
| White-collar high-skilled | 62.20 | 61.46 | 64.28 | 61.65 | 69.81 | 61.79 |
| White-collar low-skilled | 10.84 | 10.10 | 10.84 | 10.10 | 8.22 | 10.42 |
| Blue-collar high-skilled | 8.76 | 8.60 | 8.74 | 8.63 | 6.96 | 8.73 |
| Blue-collar low-skilled | 0.84 | 0.72 | 0.76 | 0.71 | 0.64 | 0.72 |
| Missing/unknown | 17.36 | 19.12 | 15.38 | 18.91 | 14.37 | 18.34 |
| **Total (%)** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **Total (n)** | **12 551** | **235 591** | **10 370** | **235 591** | **5 475** | **235 591*** ** |

Notes: * Population distribution (weights in all waves should approximate this distribution) – column shaded.

** 2009 has been chosen as it is the latest wave of data and as such is the most affected by attrition. However, any wave of data could be presented.

*** The sum of the weights in each wave sums back to the original population totals. In undertaking the weighting, no regard is given to death, or immigration over time. The population of interest is the number of 15-year-olds who were in school in 2003.

**Table 6 Y03 weights – achievement scores, 2003 and 2009**

| | PISA 2003 | | LSAY Y03 2003 | | LSAY Y03 2009 | |
|---|---|---|---|---|---|---|
| Plausible value | Unweighted mean | Population* | Unweighted mean | Weighted mean | Unweighted mean | Weighted mean |
| PV1MATH | 522.40 | 524.08 | 528.71 | 524.46 | 553.15 | 523.03 |
| PV1READ | 524.14 | 525.67 | 531.20 | 526.06 | 554.44 | 524.54 |
| PV1PROB | 528.65 | 529.91 | 534.87 | 530.25 | 557.18 | 528.90 |
| PV1SCIE | 523.10 | 525.38 | 529.72 | 525.73 | 555.41 | 524.02 |

Note: * Population distribution (weights in all waves should approximate this distribution) – column shaded.

## Y06 weights

Table 7 presents the effects of the weights for the Y06 cohort in the 2006 and 2009 waves of data.

**Table 7 Y06 weights, 2006 and 2009**

| | PISA 2006 | | LSAY Y06 2009 | |
|---|---|---|---|---|
| | Unweighted % | Population* % | Unweighted % | Final weight % |
| **State (STATE)** | | | | |
| ACT | 6.96 | 2.03 | 7.59 | 2.01 |
| NSW | 23.80 | 32.62 | 23.62 | 32.70 |
| Vic. | 16.03 | 23.96 | 16.93 | 23.96 |
| QLD | 16.95 | 19.63 | 16.72 | 19.66 |
| SA | 11.24 | 8.07 | 12.33 | 8.10 |
| WA | 10.47 | 10.23 | 10.63 | 10.16 |
| Tas. | 9.10 | 2.63 | 8.45 | 2.62 |
| NT | 5.44 | 0.82 | 3.71 | 0.79 |
| **Sector (SECTOR)** | | | | |
| Government | 60.97 | 61.68 | 56.76 | 61.61 |
| Catholic | 22.55 | 22.08 | 24.51 | 22.17 |
| Independent | 16.47 | 16.23 | 18.73 | 16.21 |
| **Geographic location of school (GEOLOC_3)** | | | | |
| Metropolitan | 67.57 | 70.47 | 70.08 | 70.59 |
| Provincial | 29.31 | 27.44 | 27.54 | 27.37 |
| Remote | 3.13 | 2.09 | 2.38 | 2.04 |
| **Indigenous status (INDIG)** | | | | |
| Non-Indigenous | 92.38 | 97.07 | 95.22 | 97.22 |
| Indigenous | 7.62 | 2.93 | 7.78 | 2.78 |
| **Highest parental education (HISCED)** | | | | |
| None | 0.93 | 0.92 | 0.62 | 1.00 |
| ISCED 1 | 0.42 | 0.36 | 0.34 | 0.32 |
| ISCED 2 | 9.84 | 9.39 | 8.66 | 9.47 |
| ISCED 3B, C | 2.99 | 2.93 | 2.41 | 3.14 |
| ISCED 3A, 4 | 31.33 | 31.32 | 29.76 | 31.45 |
| ISCED 5B | 13.97 | 14.43 | 13.71 | 14.32 |
| ISCED 5A, 6 | 37.69 | 38.20 | 43.25 | 38.23 |
| NA | 0.49 | 0.43 | 0.00 | 0.00 |
| Invalid | 0.09 | 0.10 | 0.04 | 0.09 |
| Missing | 2.24 | 1.94 | 1.21 | 1.96 |
| **Grade (ST01Q01)** | | | | |
| Year 8 | 0.10 | 0.09 | 0.05 | 0.11 |
| Year 9 | 9.37 | 9.22 | 8.04 | 9.09 |
| Year 10 | 71.65 | 70.79 | 72.01 | 70.75 |
| Year 11 | 18.79 | 19.84 | 19.84 | 19.20 |
| Year 12 | 0.08 | 0.06 | 0.05 | 0.06 |
| **Gender (ST04Q01)** | | | | |
| Male | 50.76 | 51.14 | 48.05 | 51.39 |
| Female | 49.24 | 48.86 | 51.95 | 48.61 |

| | PISA 2006 | | LSAY Y06 2009 | |
|---|---|---|---|---|
| | Unweighted % | Population* % | Unweighted % | Final weight % |
| **Immigration status (AUSIMMIG)** | | | | |
| Australian-born | 60.22 | 58.77 | 60.04 | 58.96 |
| First generation students | 29.10 | 29.81 | 30.80 | 30.12 |
| Foreign-born students | 8.25 | 9.13 | 7.74 | 9.11 |
| NA | 0.50 | 0.43 | 0.03 | 0.02 |
| Invalid | 0.11 | 0.09 | 0.03 | 0.14 |
| Missing | 1.82 | 1.76 | 1.37 | 1.66 |
| **Maths achievement quartile (PV1math)** | | | | |
| 1st quartile | 25.05 | 23.51 | 15.82 | 23.37 |
| 2nd quartile | 25.08 | 25.33 | 23.67 | 25.43 |
| 3rd quartile | 24.91 | 25.35 | 27.72 | 25.32 |
| 4th quartile | 24.96 | 25.81 | 32.79 | 25.87 |
| **Total (%)** | **100.00** | **100.00** | **100.00** | **100.00** |
| **Total (n)** | **14 170** | **234 940** | **7 299** | **234 940**** |

Notes: * Population distribution (weights in all waves should approximate this distribution) – column shaded.

** The sum of the weights in each wave sums back to the original population totals. In undertaking the weighting, no regard is given to death, or immigration over time. The population of interest is the number of 15-year-olds who were in school in 2006.

## Distribution of weights – Y03 and Y06

Tables 8 and 9 present the distributions of the LSAY weights (and normalised weights) for all years of the Y03 and Y06 cohorts.

**Table 8 Summary statistics for Y03 weights**

| Year | Variable | Mean | Sum | Sample size (n) | Minimum | Maximum |
|---|---|---|---|---|---|---|
| **Weights** | | | | | | |
| PISA | W_FSTUWT | 18.77 | 235 591 | 12 551 | 1.27 | 227.38 |
| 2003 | WT2003_P | 22.72 | 235 591 | 10 370 | 1.51 | 352.69 |
| 2004 | WT2004_P | 25.12 | 235 591 | 9 378 | 1.61 | 404.29 |
| 2005 | WT2005_P | 27.11 | 235 591 | 8 691 | 1.67 | 417.70 |
| 2006 | WT2006_P | 30.51 | 235 591 | 7 721 | 1.78 | 482.58 |
| 2007 | WT2007_P | 35.38 | 235 591 | 6 658 | 1.99 | 568.67 |
| 2008 | WT2008_P | 38.79 | 235 591 | 6 074 | 2.07 | 606.17 |
| 2009 | WT2009_P | 43.03 | 235 591 | 5 475 | 2.14 | 696.94 |
| **Normalised weights** | | | | | | |
| 2003 | WT2003 | 1.00 | 10 370 | 10 370 | 0.07 | 15.52 |
| 2004 | WT2004 | 1.00 | 9 378 | 9 378 | 0.06 | 16.09 |
| 2005 | WT2005 | 1.00 | 8 691 | 8 691 | 0.06 | 15.41 |
| 2006 | WT2006 | 1.00 | 7 721 | 7 721 | 0.06 | 15.82 |
| 2007 | WT2007 | 1.00 | 6 658 | 6 658 | 0.06 | 16.07 |
| 2008 | WT2008 | 1.00 | 6 074 | 6 074 | 0.05 | 15.63 |
| 2009 | WT2009 | 1.00 | 5 475 | 5 475 | 0.05 | 16.20 |

**Table 9 Summary statistics for Y06 weights**

| Year | Variable | Mean | Sum | Sample size (n) | Minimum | Maximum |
|---|---|---|---|---|---|---|
| **Weights** | | | | | | |
| 2006 (PISA) | W_FSTUWT | 16.58 | 234 939 | 14 170 | 1.05 | 70.74 |
| 2007 | WT2007_P | 25.12 | 234 939 | 9 353 | 1.40 | 565.49 |
| 2008 | WT2008_P | 28.04 | 234 939 | 8 380 | 1.52 | 427.67 |
| 2009 | WT2009_P | 32.19 | 234 939 | 7 299 | 1.62 | 458.05 |
| **Normalised weights** | | | | | | |
| 2006 | WT2006 | 1.00 | 14 170 | 14 170 | 0.06 | 4.27 |
| 2007 | WT2007 | 1.00 | 9 353 | 9 353 | 0.06 | 22.51 |
| 2008 | WT2008 | 1.00 | 8 380 | 8 380 | 0.05 | 15.25 |
| 2009 | WT2009 | 1.00 | 7 299 | 7 299 | 0.05 | 14.23 |

The mean columns in tables 8 and 9 show the number of individuals in the population that a single survey respondent represents. For example, in table 9, a single respondent in the 2006 PISA survey represents 16.58 others who have background characteristics similar to themselves. As attrition in LSAY over the waves increases, each individual respondent represents a larger and larger number of similar peers. The sum column shows the total returned when applying these weights to the analysis.

# Recommendations for applying weights

- Most statistical analysis software allows for the use of weights in analysis. In particular, many programs have specialised routines for the analysis of survey data (such as surveylogistic, surveyreg in SAS, svy in STATA and the survey package in R). There is no reason not to apply weights when analysing survey data. Weights should always be applied when determining means, quintiles and other such measures of a population.
- When using the LSAY data, the most appropriate weight is the final weight. This weight has been created to account for both the sampling scheme employed and the effects of attrition. For producing summary measures on any given wave of the data, this is the weight to use.
- However, in some techniques, the use of weights can result in the mis-specification of standard errors, significance tests and other relevant parameters. This arises because the estimation of these parameters is based on the weighted N, rather than the actual n (sample size). When this arises, users should use the normalised weights (weights that sum to the sample size) rather than those that return population totals. These weights are included in the LSAY datasets.
- In the case of more complicated data analysis, such as the use of mixed models, there is no clear approach to the use of weights. The choice depends on the nature of the problem and how researchers plan on reporting and using the results (for example, reporting associations that might exist in the entire population or simple associations that are seen in this dataset). An alternative approach to directly applying weights is to include all the variables used to create the weights as independent variables. This will result in an unbiased estimate, correct standard errors and inferences. However, under-specifying (failing to include certain variables) the model can lead to biased estimates and incorrect standard errors. Conversely, over-specifying models by including too many covariates can lead to estimation problems, and so researchers need to consider their models carefully and undertake appropriate diagnostics. It is important for researchers to note that the use of weights typically has a much more pronounced effect on descriptive statistics than on regression coefficients.

Users of the LSAY data undertaking complex analytical techniques should be aware of the different ways that weights can be used in their analysis. The ABS (2008) provides a comprehensive list of references on the use of survey weights in modelling.


- Chambers, RL & Skinner, CJ (eds) 2003, *Analysis of survey data*, Wiley, Chichester, England.
- DuMouchel, WH & Duncan GJ 1983, 'Using sample survey weights in multiple regression analyses of stratified samples', *Journal of the American Statistical Association*, vol.78, no.383, pp.535—43.
- Magee, L, Robb, AL & Burbidge, JB 1998, 'On the use of sampling weights when estimating regression models with survey data', *Journal of Econometrics*, vol.84, issue 2, pp.251—71.
- Pfeffermann, D 1993, 'The role of sampling weights when modeling survey data', *International Statistical Review*, vol.61, no.2, pp.317—37.

- Pfeffermann, D 1996, 'The use of sampling weights for survey data analysis', *Statistical Methods in Medical Research*, 5, pp.239—61.
- Skinner, CJ, Holt, D & Smith, TMF 1989, *Analysis of complex surveys*, Wiley, Chichester, England.
- Winship, C & Radbill, L 1994, 'Sampling weights and regression analysis', *Sociological Methods and Research*, vol.23, no.2, pp.230—57.

# References


ABS (Australian Bureau of Statistics) 2008, 'Frequently asked questions — tips for using CURFs: how should I use survey weights in my model?', ABS, Canberra, viewed July 2011, <http://www.abs.gov.au/websitedbs/d3310114.nsf/4a256353001af3ed4b2562bb00121564/d4b021fd647c719cca257362001e13dc!OpenDocument>.

NCVER (National Centre for Vocational Education Research) 2011a, *Longitudinal Surveys of Australian Youth (LSAY) 2003 cohort: user guide*, Technical report 54, NCVER, Adelaide.

——2011b, *Longitudinal Surveys of Australian Youth (LSAY) 2006 cohort: user guide*, Technical report 55, NCVER, Adelaide.

OECD (Organisation for Economic Co-operation and Development) 2005, 'PISA 2003 technical report', OECD, Paris, viewed July 2011, <http://www.oecd.org/dataoecd/49/60/35188570.pdf>.

——2009, 'PISA 2006 technical report', OECD, Paris, viewed July 2011, <http://www.oecd.org/dataoecd/0/47/42025182.pdf>.

Rothman, S 2007, *Sampling and weighting of the 2003 LSAY cohort*, Technical report 43, Australian Council for Educational Research, Camberwell, Vic.

Thomson, S, Creswell, J & De Bortoli, L 2004, *Facing the future: a focus on mathematical literacy among Australian 15-year-old students in PISA 2003*, Australian Council for Educational Research, Camberwell, Vic.

Thomson, S & De Bortoli, L 2008, *Exploring scientific literacy: how Australia measures up*, Australian Council for Educational Research, Camberwell, Vic.

# Appendix A: Sample weights

The variables used to calculate the sample weights are state and school.

## Y03 Sample weights

Table A1 shows the raw and weighted percentages for the 2003 PISA cohort. This table clearly shows the impact of applying sample weights to the data. The distribution of the population variables is shown in column 2, and it can be seen that, with the weights applied, the distributions of the 2003 and 2009 data match the PISA distributions. It is also clear that some jurisdictions are over-sampled and some are under-sampled. In particular, the Australian Capital Territory comprises 7% of the raw data, yet the true population proportion is around 2%. Conversely, New South Wales is under-sampled, with 24% of the raw data, but 32% of the population.

**Table A1 Y03 sample weights, 2003 and 2009**

| | PISA 2003 | | LSAY Y03 2003 | | LSAY Y03 2009** | |
|---|---|---|---|---|---|---|
| | Unweighted % | Population* % | Unweighted % | Sample weight % | Unweighted % | Sample weight % |
| **State** | | | | | | |
| ACT | 7.12 | 1.89 | 7.00 | 1.89 | 7.82 | 1.88 |
| NSW | 23.78 | 31.65 | 22.79 | 31.67 | 22.89 | 31.85 |
| Vic. | 18.76 | 24.13 | 19.48 | 24.12 | 20.02 | 24.17 |
| QLD | 15.41 | 19.26 | 15.73 | 19.26 | 15.49 | 19.13 |
| SA | 9.83 | 8.95 | 10.02 | 8.95 | 10.06 | 8.88 |
| WA | 14.08 | 11.12 | 14.29 | 11.10 | 13.44 | 11.09 |
| Tas. | 6.41 | 2.25 | 6.56 | 2.25 | 6.65 | 2.25 |
| NT | 4.65 | 0.75 | 4.14 | 0.75 | 3.63 | 0.75 |
| **Sector** | | | | | | |
| Government | 64.58 | 61.84 | 64.06 | 61.85 | 60.00 | 61.86 |
| Catholic | 19.62 | 21.14 | 20.59 | 21.12 | 21.79 | 21.15 |
| Independent | 15.79 | 17.03 | 15.35 | 17.04 | 18.21 | 16.99 |
| **Total (%)** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **Total (n)** | **12 551** | **235 591** | **10 370** | **235 591** | **5 475** | **235 591*** ** |

Notes: * Population distribution (weights in all waves should approximate this distribution) – column shaded.

** 2009 has been chosen as it is the latest wave of data and as such is the most affected by attrition. However, any wave of data could be presented.

*** The sum of the weights in each wave returns to the original population totals. In calculating the weights, no regard is given to death, or immigration over time. The population of interest is the number of 15-year-olds who were in school in 2003.

# Y06 sample weights

Table A2 presents the raw and weighted percentages for the 2006 and 2009 waves of data.

**Table A2 Y06 sample weights, 2006 and 2009**

| | PISA 2006 | | LSAY Y06 2009** | |
|---|---|---|---|---|
| | Unweighted % | Population* % | Unweighted % | Sample weight % |
| **State** | | | | |
| ACT | 6.96 | 2.03 | 7.59 | 2.04 |
| NSW | 23.80 | 32.62 | 23.62 | 32.65 |
| Vic. | 16.03 | 23.96 | 16.93 | 23.98 |
| QLD | 16.95 | 19.63 | 16.72 | 19.65 |
| SA | 11.24 | 8.07 | 12.33 | 8.08 |
| WA | 10.47 | 10.23 | 10.63 | 10.18 |
| Tas. | 9.10 | 2.63 | 8.45 | 2.63 |
| NT | 5.44 | 0.82 | 3.71 | 0.80 |
| **Sector** | | | | |
| Government | 60.97 | 61.68 | 56.76 | 61.73 |
| Catholic | 22.55 | 22.08 | 24.51 | 22.10 |
| Independent | 16.47 | 16.23 | 18.73 | 16.16 |
| **Total (%)** | **100.00** | **100.00** | **100.00** | **100.00** |
| **Total (n)** | **14 170** | **234 940** | **7 299** | **234 940** |

Notes: * Population distribution (weights in all waves should approximate this distribution) – column shaded.

** 2009 has been chosen as it is the latest wave of data and as such is the most affected by attrition. However, any wave of data could be presented.

# Appendix B: Attrition weights

The variables used to calculate the attrition weights do not include those used in the sampling scheme. The variables included are those that are most likely to be related to the chance that an individual will not respond to the survey. For example, those who are the least academically successful, from low socioeconomic households, immigrants, or Indigenous people are less likely to respond. Young people are highly mobile and an aspect of the attrition observed in LSAY is due to young people moving around, going overseas and other such factors.

For the 2003 cohort, variables included in the weights for attrition must be those that were included in the PISA questionnaire; for this reason Indigenous status is not included as this is an LSAY variable.

The following sections outline the effects of applying the attrition weights for selected years of the 2003 and 2006 cohorts.

## Y03 attrition weights

Table B1 shows the effects of the raw and weighted percentages for attrition variables in Y03. The table demonstrates that the use of attrition weights ensures that the distributions of the variables in the reduced samples match those observed in the original PISA population.

**Table B1 Y03 attrition weights, 2003 and 2009**

| | PISA 2003 | | LSAY Y03 2003 | | LSAY Y03 2009** | |
|---|---|---|---|---|---|---|
| | Unweighted | Population* | Unweighted | Attrition weight | Unweighted | Attrition weight |
| | % | % | % | % | % | % |
| **Family structure (FAMSTRUC)** | | | | | | |
| Single parent family | 21.17 | 19.80 | 19.97 | 19.84 | 16.24 | 20.17 |
| Nuclear family | 66.72 | 68.69 | 68.94 | 68.87 | 74.41 | 68.69 |
| Mixed family | 8.10 | 7.79 | 8.00 | 7.77 | 6.74 | 7.69 |
| Other | 2.88 | 2.70 | 2.47 | 2.62 | 2.03 | 2.57 |
| Missing/unknown | 1.13 | 1.03 | 0.62 | 0.90 | 0.58 | 0.89 |
| **Highest educational level of parents (HISCED)** | | | | | | |
| None | 1.25 | 1.38 | 1.21 | 1.39 | 1.02 | 1.36 |
| ISCED 1 | 0.50 | 0.56 | 0.39 | 0.51 | 0.29 | 0.58 |
| ISCED 2 | 11.19 | 10.97 | 10.98 | 11.00 | 8.91 | 11.11 |
| ISCED 3B, C | 2.36 | 2.22 | 2.36 | 2.23 | 1.55 | 2.28 |
| ISCED 3A, 4 | 29.04 | 29.32 | 29.46 | 29.27 | 27.18 | 29.46 |
| ISCED 5B | 13.42 | 13.87 | 13.55 | 13.88 | 13.11 | 13.67 |
| ISCED 5A, 6 | 39.21 | 38.77 | 39.75 | 38.89 | 46.43 | 38.68 |
| Missing/Unknown | 3.04 | 2.93 | 2.30 | 2.84 | 1.60 | 2.86 |
| **Grade (relative to modal grade, Year 10)** | | | | | | |
| -3 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |
| -2 | 0.13 | 0.14 | 0.11 | 0.14 | 0.05 | 0.20 |
| -1 | 8.48 | 8.34 | 8.37 | 8.31 | 8.33 | 8.56 |
| 0 | 71.56 | 72.26 | 71.15 | 72.29 | 71.14 | 71.79 |
| 1 | 19.76 | 19.21 | 20.30 | 19.20 | 20.38 | 19.40 |
| 2 | 0.06 | 0.05 | 0.07 | 0.05 | 0.09 | 0.05 |
| **Sex (ST03Q01)** | | | | | | |
| Male | 49.53 | 49.17 | 50.39 | 49.18 | 49.97 | 49.65 |
| Female | 50.47 | 50.83 | 49.61 | 50.82 | 50.03 | 50.35 |
| **Immigration status (IMMIG)** | | | | | | |
| Native students | 77.14 | 75.54 | 78.42 | 75.58 | 78.10 | 75.18 |
| First generation students | 10.69 | 11.48 | 10.65 | 11.60 | 10.76 | 11.56 |
| Non-native students | 10.02 | 10.75 | 9.40 | 10.75 | 9.88 | 11.09 |
| Missing/unknown | 2.14 | 2.24 | 1.53 | 2.08 | 1.26 | 2.17 |
| **Occupational status of parents (SSECATEG)** | | | | | | |
| White-collar high-skilled | 62.20 | 61.46 | 64.28 | 61.65 | 69.81 | 60.80 |
| White-collar low-skilled | 10.84 | 10.10 | 10.84 | 10.11 | 8.22 | 10.42 |
| Blue-collar high-skilled | 8.76 | 8.60 | 8.74 | 8.62 | 6.96 | 8.68 |
| Blue-collar low-skilled | 0.84 | 0.72 | 0.76 | 0.71 | 0.64 | 0.73 |
| Missing/unknown | 17.36 | 19.12 | 15.38 | 18.90 | 14.37 | 18.36 |
| **Total (%)** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **Total (n)** | **12 551** | **235 591** | **10 370** | **235 591** | **5 475** | **235 591** |

Notes: * Population distribution (weights in all waves should approximate this distribution) – column shaded.

** 2009 has been chosen as it is the latest wave of data and as such is the most affected by attrition. However, any wave of data could be presented.

Table B2 shows the effect of attrition weights on the continuous achievement variables. It is clear that the use of weights corrects an upward bias caused by attrition.

**Table B2 Y03 attrition weights – achievement scores, 2003 and 2009**

| | PISA 2003 | | LSAY Y03 2003 | | LSAY Y03 2009 | |
|---|---|---|---|---|---|---|
| Plausible value* | Unweighted mean | Population** | Unweighted mean | Attrition weighted mean | Unweighted mean | Attrition weighted mean |
| PV1MATH | 522.40 | 524.08 | 528.71 | 524.50 | 553.15 | 523.17 |
| PV1READ | 524.14 | 525.67 | 531.20 | 526.11 | 554.44 | 524.75 |
| PV1PROB | 528.65 | 529.91 | 534.87 | 530.29 | 557.18 | 529.09 |
| PV1SCIE | 523.10 | 525.38 | 529.72 | 525.80 | 555.41 | 524.29 |

Notes: * Only plausible values are shown.

** Population distribution (weights in all waves should approximate this distribution) – shaded column.

# Y06 attrition weights

Table B3 presents the distribution of attrition variables for 2006 and 2009. We observe that the use of attrition weights corrects the effects of attrition and ensures that the distribution of these variables matches those observed in the original PISA population.

**Table B3 Y06 attrition weights, 2006 and 2009**

| | PISA 2006 | | LSAY Y06 2009** | |
|---|---|---|---|---|
| | Unweighted | Population* | Unweighted | Attrition weight |
| | % | % | % | % |
| **Geographic location of school (GEOLOC_3)** | | | | |
| Metropolitan | 67.57 | 70.47 | 70.08 | 70.45 |
| Provincial | 29.31 | 27.44 | 27.54 | 27.47 |
| Remote | 3.13 | 2.09 | 2.38 | 2.08 |
| **Indigenous status (INDIG)** | | | | |
| Non-Indigenous | 92.38 | 97.07 | 95.22 | 97.18 |
| Indigenous | 7.62 | 2.93 | 4.78 | 2.8 |
| **Highest parental education (HISCED)** | | | | |
| None | 0.93 | 0.92 | 0.62 | 0.95 |
| ISCED 1 | 0.42 | 0.36 | 0.34 | 0.35 |
| ISCED 2 | 9.84 | 9.39 | 8.66 | 9.46 |
| ISCED 3B, C | 2.99 | 2.93 | 2.41 | 3.03 |
| ISCED 3A, 4 | 31.33 | 31.32 | 29.76 | 31.52 |
| ISCED 5B | 13.97 | 14.43 | 13.71 | 14.48 |
| ISCED 5A, 6 | 37.69 | 38.20 | 43.25 | 38.20 |
| NA | 0.49 | 0.43 | 0.00 | 0.00 |
| Invalid | 0.09 | 0.10 | 0.04 | 0.10 |
| Missing | 2.24 | 1.94 | 1.21 | 1.91 |
| **Grade (ST01Q01)** | | | | |
| Year 8 | 0.10 | 0.09 | 0.05 | 0.11 |
| Year 9 | 9.37 | 9.22 | 8.04 | 9.16 |
| Year 10 | 71.65 | 70.79 | 72.01 | 70.82 |
| Year 11 | 18.79 | 19.84 | 19.84 | 19.84 |
| Year 12 | 0.08 | 0.06 | 0.05 | 0.05 |
| **Gender (ST04Q01)** | | | | |
| Male | 50.76 | 51.14 | 48.05 | 51.24 |
| Female | 49.24 | 48.86 | 51.95 | 48.76 |
| **Immigration status (AUSIMMIG)** | | | | |
| Australian-born | 60.22 | 58.77 | 60.34 | 59.02 |
| First generation students | 29.10 | 29.81 | 30.80 | 30.11 |
| Foreign-born students | 8.25 | 9.13 | 7.74 | 9.10 |
| NA | 0.50 | 0.43 | 0.03 | 0.02 |
| Invalid | 0.11 | 0.09 | 0.03 | 0.12 |
| Missing | 1.82 | 1.76 | 1.37 | 1.63 |
| **Maths achievement quartile (PV1math)** | | | | |
| 1st quartile | 25.05 | 23.51 | 15.82 | 23.33 |
| 2nd quartile | 25.08 | 25.33 | 23.67 | 25.36 |
| 3rd quartile | 24.91 | 25.35 | 27.72 | 25.38 |
| 4th quartile | 24.96 | 25.81 | 32.79 | 25.94 |
| **Total (%)** | **100.00** | **100.00** | **100.00** | **100.00** |
| **Total (n)** | **14 170** | **234 940** | **7 299** | **234 940** |

Notes: * Population distribution (weights in all waves should approximate this distribution).

** 2009 has been chosen as it is the latest wave of data and as such is the most affected by attrition. However, any wave of data could be presented.


**Australian Government**

**Department of Education, Employment and Workplace Relations**


National Centre for Vocational Education Research Ltd
Level 11, 33 King William Street, Adelaide, South Australia
PO Box 8288, Station Arcade, SA 5000 Australia
Telephone +61 8 8230 8400 Facsimile +61 8 8212 3436
Website www.ncver.edu.au Email ncver@ncver.edu.au