SUPPORT DOCUMENT

School completion: data appendix

Jacqueline Homel
Astghik Mavisakalyan
Ha Trong Nguyen
Chris Ryan

Social Policy Evaluation, Analysis and Research Centre
Australian National University

This document was produced by the author(s) based on their research for the report *School completion: what we learn from different measures of family background*, and is an added resource for further information. The report is available on NCVER's website: <http://www.ncver.edu.au>.

Contents

Tables and figures

Tables

Figures

Data appendix

This study relies on two main sources to analyse how Year 12 completion is related to disadvantage: the Longitudinal Surveys of Australian Youth (LSAY) and the Youth in Focus (YIF) Project. This appendix provides details on the data used. First, a detailed description of all variables controlled in the regression is provided. These include variables overlapping in LSAY and YIF as well as those unique to each of the sources. Second, the procedure of generating weights in LSAY is described.

Variable definition

Detailed description of variables used in the prediction of year 12 completion is provided below. Variables shared across LSAY and YIF as well as unique to each of the sources are described. Details in measurement periods are shown separately for sources where differences exist.

The dependent variable, year 12 completion, is defined to equal to 1 if the respondents report having a year 12 certificate. Young people still at school in the observed period are excluded from the sample.

Variable name	Description	Data source	Measurement period
Overlapping items in LSAY and YIF			
Gender	=1 if male, =0 if female	LSAY 2003 (wave 1) telephone survey	respondent age 15
		YIF, youth's questionnaire	time of survey
Lives in major city	=1 if living in major city of Australia, =0 otherwise	LSAY 2003 (wave 1) telephone survey	respondent's last school year
		YIF, youth's questionnaire	time of survey
Indigenous Australian	= 1 if Aboriginal or Torres Strait Islander, =0 otherwise	LSAY, PISA student questionnaire	respondent age 15
		YIF, youth's questionnaire	time of survey
Father or head-of-household employed	= 1 if father employed full-time or part-time, = 0 if not employed.	LSAY, PISA student questionnaire	respondent age 15
		YIF, youth's questionnaire	respondent age 14
Parental education	=1 if father (mother) has Bachelor degree or higher, =0 if less	LSAY, PISA student questionnaire	respondent age 15
		YIF, parent's questionnaire	respondent age 14
Number of siblings	number of siblings (4=4 or more in LSAY)	LSAY, PISA student questionnaire	respondent age 15
		YIF, parent's questionnaire	time of survey
Respondent's country of birth	Three dummy variables indicating students who were born in (1) Australia (base group), (2) overseas English-speaking country (UK, Ireland, Canada, NZ, South Africa, USA), (3) overseas non-English speaking country	LSAY, PISA student questionnaire	respondent age 15
		YIF, youth's questionnaire	time of survey

Variable name	Description	Data source	Measurement period
Two-parent household	=1 if both parents live in household, =0 if single parent household	LSAY, PISA student questionnaire YIF, youth's questionnaire	respondent age 15 respondent age 14
School sector	Three dummy variables indicating students enrolled in (1) Government (base group), (2) Catholic, (3) Independent schools	PISA student questionnaire for 2003, LSAY surveys, waves 2-4 for 2004-2006 YIF, youth's questionnaire	respondent's last school year last school the respondent attended
Residential state	Dummies for the state of residence of the respondent. NSW (base), VIC, QLD, SA, WA, TAS, NT	LSAY, PISA student questionnaire	respondent age 15
	Dummies for the state of residence of the respondent. NSW (base), VIC, QLD, SA, WA or NT, TAS	YIF, youth's questionnaire	time of survey
SEIFA scales (multiplied by 1000):	see below for details	LSAY 2003 (wave 1) telephone survey (residential postcode) YIF, youth's questionnaire	respondent age 15 time of survey
SES: responding parent ANU 4 (multiplied by 1000)	Occupational status of the responding parent of the youth measured by the ANU4 scale (Jones and McMillan 2001)	PISA student questionnaire YIF, parent's questionnaire	respondent age 15 time of survey
Self-assessed school performance	Four dummy variables indicating students who are doing (1) well above average (base), (2) better than average, (3) average, and (4) below average.	LSAY wave 2003 (wave 1) telephone survey YIF, youth's questionnaire	respondent age 15 last year at school
Items unique to LSAY			
Achievement (all scores multiplied by 1,000)			
Reading literacy	see below	LSAY, PISA student questionnaire	respondent age 15
(reading literacy if above average)			
Mathematical literacy			
(mathematical literacy if above average)			
Aspirations			
Expected job at age 30 (ANU4) (multiplied by 1,000)	Job that respondent expects to have at the age of 30- ANU4 scale (Jones and McMillan 2001)	LSAY, PISA student questionnaire	respondent age 15
Plans to participate in year 12	=1 if student plans to go on to year 12, =0 if not	LSAY wave 2003 (wave 1) telephone survey	respondent age 15
Plans to go to university	=1 if student plans to go to university, =0 if otherwise		
Ever repeated a school grade	If the respondent has ever repeated a school grade.	LSAY, PISA student questionnaire	respondent age 15
School climate			
Student-body attitudes to school	Principals' reports of overall positive attitudes of students in their school towards learning and behaviour. Higher scores indicate more positive attitudes.	LSAY, PISA school questionnaire	respondent age 15

Variable name	Description	Data source	Measurement period
Teachers' attitudes to school	Principals' reports of overall morale and commitment of teachers in their school. Higher scores indicate more positive attitudes.		
Teacher problem behaviour	Principals' reports of the extent to which teacher behaviours like low expectations of students, absenteeism, and poor teacher-student relations hindered the learning of students in their school. Higher scores indicate higher levels of problem behaviour.		
Student-body problem behaviour	Principals' reports of the extent to which student behaviours like truancy, use of substances, and bullying hindered the learning of students in their school. Higher scores indicate higher levels of problem behaviour.		
School characteristics			
School average SES (ANU4)	Head of household ANU4 scores averaged across students in schools	PISA student questionnaire	respondent age 15
School size	Number of students in the school	PISA school questionnaire	respondent age 15
Proportion of girls to boys	Proportion of girls to boys in the school		
Grade structure	Grades present in the students high school: 1- high school grades only (the base category- grades 8-12 in QLD and WA, and 7-12 in other states), 2- grades K-12, 3- senior high school grades only (11-12).		
School selectivity	Two dummy variables indicating selection based on student (1) residence in a particular area (2) academic record		
Extent of maths streaming within school	Whether the school organises instruction differently for students with different maths abilities. 1= not for any classes, 2= for some classes, 3= for all classes		
Estimated number of student assessments per year	Principals' reports of how often each year students are assessed using a variety of methods. 1=<20. 2= 20-39, 3= >40		
School resources			
Student-teacher ratio	Student-teacher ratio	PISA school questionnaire	respondent age 15
Adequacy of physical facilities	Principals' reports of the extent to which their school's capacity to provide instruction is hindered by e.g. building quality, classroom space etc. Higher scores indicate better resources.		
Adequacy of school educational resources	Principals' reports of the extent to which their school's capacity to provide instruction is hindered by e.g. lack of instructional materials, computers, science equipment. Higher scores indicate better resources.		
Extent of teacher shortage	Principals' reports of teacher shortages across a range of subject areas.		
Items unique to YIF			
Family income			
Family income-parental income (AU$1,000,000)	Total family income of principal responding parent (and partner) in 1000 AU$	YIF, parent's questionnaire	time of survey
Family income is zero or not able to be estimated	=1 if no income or can't say, = 0 if have or refusal	YIF, parent's questionnaire	time of survey

Variable name	Description	Data source	Measurement period
Family income support	Four dummies indicating family's welfare receipt history (1) no history of welfare receipt (non-recipient, the base group) (2) intensive income support: the family received welfare for more than six years while the youth growing up (3) late moderate income support: family received welfare for less than six years after 1998 (4) early moderate income support: family received welfare for less than six years before 1998	YIF, youth's questionnaire	time of survey
Number of marriages/relationships principal responding parent	Number of marriages/relationships principal responding parent had	YIF, parent's questionnaire	time of survey
Parent factors			
Parents often read to respondent at night when younger	= 1 if parents often read at night when young, = 0 otherwise	YIF, youth's questionnaire	time of survey
Parents often helped respondent with schoolwork	= 1 if parents often help with homework, = 0 otherwise	YIF, youth's questionnaire	time of survey
Respondent's student experiences			
Ever repeated school	= 1 if ever repeated school, = 0 if not	YIF, youth's questionnaire	time of survey
Number of primary and secondary schools attended	Number of primary and secondary schools the youth has attended	YIF, youth's questionnaire	time of survey
Ever suspended from secondary school	= 1 if ever been suspended from secondary school	YIF, youth's questionnaire	time of survey
Risky behaviours			
Ever smoked	= 1 if ever smoked	YIF, youth's questionnaire	time of survey
Ever drunk alcohol	=1 if ever drunk alcohol	YIF, youth's questionnaire	time of survey
Ever tried marijuana	=1 if ever tried marijuana	YIF, youth's questionnaire	time of survey

SEIFA scales

SEIFA stands for Socio-economic Index for Areas. The SEIFA indexes are rankings of socio-economic characteristics for geographic areas (Census collection districts, or CDs) across Australia. Rankings are created by the ABS based on characteristics of people, families and dwellings drawn from the five-yearly Census. There are four indexes in total, and these rank areas in terms of their 'Disadvantage'; their relative 'Advantage and Disadvantage'; their access to 'Economic Resources'; and to the average 'Education and Occupation' skills of the people living in an area. This report uses the relative 'Advantage and Disadvantage' scale and the 'Education and Occupation' scale.

The Advantage/Disadvantage scale is based on 21 measures, including low or high income, internet connection and occupation and education. A higher score reflects relative advantage (ABS intro). The Education/Occupation scale is based on measures of education achieved and/or currently undertaken, as well as the degree to which residents' occupations require high or low levels of skills. It does not include income. Higher scores reflect relatively higher education and occupation status of people in that CD. The scales are standardised such that the mean score is 1,000 and about 2/3 of the scores fall between 900 and 1,100.

The indexes are available on the ABS website, cat. no. 2033.0.55.001 Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia -Data only, 2006: http://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001

Mathematical literacy and reading literacy

The PISA survey assesses 15 year olds' mathematical literacy and reading literacy, which represent broad maths and reading concepts and skills. The emphasis is not so much on testing knowledge of specific aspects taught in the school curricula, but on mastery of processes, the understanding of concepts and the ability to function in various real-world situations within each domain. Such abilities are considered necessary to enable students to apply maths and reading skills in adult life (OECD, 2003b).

Scores on these scales are reported as plausible values. These are a set of random values (five for each scale) selected for each student at random from an estimated ability distribution of students with similar item response patterns and backgrounds. PVs are used because the goal of the PISA assessment is to provide good estimates of parameters of student populations, rather than estimates of individual student performance.

Mathematical literacy

Mathematical literacy includes students' ability to "analyse, reason, and communicate ideas effectively as they pose, formulate, solve, and interpret solutions to mathematical problems in a variety of situations." (OECD, 2003b, p. 15). It includes assessments of:

- mathematical content (quantity, space and shape, change and relationships, uncertainty)
- process (e.g. the use of mathematical language, modelling and problem solving skills)
- situations in which maths is used (including personal, educational, occupational, public and scientific)

Reading literacy

Reading literacy includes students' ability to "understand, use and reflect on written text to achieve their purposes." (OECD, 2003b, p.16). This includes an 'active' element-a capacity to reflect upon written text, as well as merely comprehend it. The reading items include assessments of:

- text format, including different prose forms (e.g. narration, argumentation) and various non-continuous texts (e.g. graphs, application forms, advertisements).

- reading process, referring to students' ability to retrieve information, understand and interpret the text, and reflect upon its content, form and features.

- the situation for which the text was constructed (e.g. a novel or letter for personal use, official documents for public use, textbooks or worksheets for educational use).

Survey weights, attrition and representativeness in Y03

For the Y03 cohort, weights are provided with the LSAY data to weight the data so that it better represents the original population from which the PISA data were drawn. The LSAY Y03 weights were designed to reflect both the original PISA weights and to undo the non-representativeness introduced into the sample from the observations lost between the PISA survey and the first LSAY survey a few months later. Some 2181 observations from the original PISA sample of 12551 were not surveyed in the first wave of LSAY.

The procedure for estimating the component of the weights to adjust for the lost observations is described in Rothman (2007). Basically it involved use of nine variables from PISA to construct 'adjustment factors' for the PISA weights. These variables included gender, location, family structure, indigenous status, parental education, intended student occupational level, student grade, education program orientation and country of birth. The exact way the adjustment factors were constructed is not spelled out, but in Table 5 of that report the author presents results that suggest the weights do a good job of reweighting the LSAY sample such that its demographic characteristics match that of the original PISA sample quite closely. This comparison is partially replicated in the first two columns of Table 1 below. The available LSAY data also contain a set of weights that are designed to deal with later attrition from the LSAY sample, so that the weighted demographic characteristics of the cohort continue to match that of the original populations from which LSAY (and hence PISA) were drawn.

The procedure to generate the weights described in Rothman (2007) departs from that used to weight the Y95 and Y98 cohorts to deal with survey attrition, described in Marks and Long (2000). There, those who remained in the survey were weighted such that the distribution of respondents in later years matched the achievement distribution observed in the original cohort. This seems important in a data set used to look extensively at education outcomes where achievement is an important determinant of those outcomes. If the sample becomes unrepresentative in terms of the achievement of individuals, it will be unrepresentative of participation in education and training of the broader population if the weights do not account for this situation.

It is clear that there are such problems in the Y03 data, in part driven by the lost 2181 observations between the PISA survey and the start of the LSAY interviews. This is demonstrated in the first three columns of Table 2 below. The Table reports the average plausible values in four learning domains in PISA: maths, reading, science and problem solving. The first column shows the reported results for Australia for the whole PISA sample (in Thomson *et al.* 2004). These are broadly in the range of 525—530.

The third column shows the average values for the group later observed in the LSAY data using the PISA weights. The average values are typically around six points higher. The second column shows the averages for the group who were not interviewed in LSAY. Their average values are 25 to 27 points lower than the PISA average, so the group not in LSAY lost from PISA had, on average, substantially lower levels of school achievement compared to those observed in LSAY. The fourth column shows how well the LSAY weights adjust the averages for the sample who remained LSAY; that is, how well the weights deal with the attrition following PISA. The weights lower the average plausible values by around 1 point, leaving the average scores around 5 points higher than those reported for the entire PISA sample.

Subsequent attrition from the Y03 cohort exacerbates the problems with the representativeness of the sample in terms of school achievement. The fifth column shows the attrition-weighted averages for the plausible values for the remaining Y03 cohort in 2006. The averages are five points higher than the weighted values for 2003 and now 10 points higher than the original PISA averages. In this dimension at least, the weights released with the data do not appear to be having their intended effect.

It is possible to adopt a different procedure to calculate weights to deal with attrition. The approach involves using the same nine variables as in the approach in Rothman (2007), but to estimate a regression equation where the dependent variable was participation or not in the LSAY survey (among those in PISA), explained by the set of explanatory variables from PISA. This set of variables can be supplement with the inclusion of other variables. Specifically, it is possible to include the plausible achievement values. The regression equation is then used to predict the probability of response for each individual in the survey and the PISA weights adjusted by multiplying them by [1/individual probability of response].

The remaining sample can then be reweighted so that the average levels of achievement match the original PISA averages much more closely. The results of this exercise are presented in the sixth and seventh columns of Table 2, which show the averages for the 2003 LSAY sample and for those who remain in the data to 2006. The achievement averages are much closer to the reported PISA averages with these weights than those provided with the data, though by 2006 the estimates with these weights are around 1 point too high.

Moreover, the adjusted weights act to make the entire distribution of achievement in the LSAY sample closer to that of the original PISA sample. This is shown with the aid of Figure 1, which shows the distribution of scores in mathematics in the original PISA sample, the distribution of those observed in LSAY in 2003, both weighted by the relevant weights provided in the data and the same LSAY distribution reweighted with the adjusted weights. The LSAY distribution with the adjusted weights matches the original PISA distribution much more closely.

The last three columns of Table 1 show how the distributions of demographic characteristics differ with the original and the adjusted weights. The third column shows the impact of attrition on the weighted distribution of demographic characteristics of those remaining in the Y03 data to 2006. The final two columns show the original 2003 distribution weighted using the adjusted weights and the 2006 distribution. The adjusted weights are more effective in matching the distribution of characteristics in the PISA sample than are the weights provided with the data.

Table 1   Distribution in 2006 on selected wave 1 variables of LSAY 2003 sample with PISA weights, LSAY weights, and revised weights

Variable	Variable categories	PISA 2003 student weight %	LSAY 2003 student weight %	LSAY 2006 attrition weight %	Revised 2003 student weight %	Revised 2006 attrition weight %
Sex	Female	49.2	49.2	48.6	49.2	49.4
	Male	50.8	50.8	51.4	50.8	50.6
State/Territory	ACT	1.9	1.9	1.9	1.9	1.9
	NSW	31.7	31.7	31.6	31.7	31.8
	VIC	24.1	24.1	24.4	24.2	24.3
	QLD	19.3	19.3	18.9	19.1	19.2
	SA	9.0	9.0	9.0	9.0	8.8
	WA	11.1	11.1	11.2	11.2	11.0
	TAS	2.2	2.2	2.2	2.2	2.2
	NT	0.7	0.7	0.7	0.7	0.7
MCEETYA Location Class	Metropolitan Zone Mainland State Capital City regions	60.6	59.7	60.4	60.7	60.5
	Metropolitan Zone Major urban Statistical Districts	11.7	12.0	12.0	11.8	11.8
	Provincial Zone Inner provincial areas	11.7	11.9	11.6	11.5	11.5
	Provincial Zone Outer provincial areas	5.7	5.9	6.1	5.7	5.7
	Provincial Zone Provincial City Statistical Districts 25,000	4.3	4.5	3.8	4.4	4.4
	Provincial Zone Provincial City Statistical Districts 50,000	5.4	5.4	5.3	5.4	5.5
	Remote Zone Remote areas	0.6	0.6	0.7	0.6	0.6
	Remote Zone Very Remote areas	0.0	0.0	0.1	0.0	0.0
Indigenous Status[1]	Non-Indigenous	98.1	98.1	98.6	97.9	97.9
	Indigenous	1.9	1.9	1.4	2.1	2.1
Family Structure*	Single parent family	20.0	18.7	18.6	20.0	20.0
	Nuclear family	69.4	71.2	71.9	69.5	69.4
	Mixed family	7.9	7.9	7.3	7.9	7.9
	Other	2.7	2.2	2.2	2.7	2.7
Highest parent White collar/Blue collar classification*	White collar high skilled	65.4	65.6	65.9	65.0	65.3
	White collar low skilled	18.4	18.5	18.2	18.8	18.6
	Blue collar high skilled	8.6	8.6	8.5	8.6	8.5
	Blue collar low skilled	7.5	7.4	7.5	7.6	7.6
Highest educational level of parents*	None – Pre-primary education	1.4	1.4	1.2	1.4	1.4
	ISCED 1 – Primary education	0.6	0.4	0.4	0.4	0.4
	ISCED 2 – Lower secondary	11.3	11.0	11.1	11.3	11.1
	ISCED 3B, C –Certificate III	2.3	2.3	2.2	2.4	2.2
	ISCED 3A, ISCED 4 – Senior secondary, Certificate IV	30.2	30.2	30.0	30.2	30.0
	ISCED 5B – Diploma level	14.3	14.7	14.4	14.5	14.7
	ISCED 5A, 6 – Bachelor degree and higher	39.9	39.9	40.6	39.8	40.1

School completion: data appendix

Variable	Variable categories	PISA 2003 student weight	LSAY 2003 student weight	LSAY 2006 attrition weight	Revised 2003 student weight	Revised 2006 attrition weight
		%	%	%	%	%
Country of birth*	Native students	77.3	77.9	78.3	77.2	77.1
	First Generation students	11.7	11.8	11.9	11.8	11.8
	Non-native students	11.0	10.3	9.9	11.0	11.1
Grade	Year 7	0.0	0.0	0.0	0.0	0.0
	Year 8	0.1	0.1	0.1	0.1	0.1
	year 9	8.3	8.4	8.3	8.7	9.6
	Year 10	72.3	72.3	71.7	71.8	71.4
	Year 11	19.2	19.1	19.9	19.3	18.8
	Year 12	0.1	0.1	0.1	0.1	0.1
Expected educational level of student (ISCED)*	None	0.4	0.3	0.3	0.3	0.4
	ISCED 2 – Lower secondary	2.7	2.3	2.3	2.6	2.3
	ISCED 3B, C –Certificate III	3.7	3.4	3.6	3.7	3.3
	ISCED 3A, ISCED 4 – Senior secondary, Certificate IV	22.7	22.6	22.6	23.0	22.7
	ISCED 5B – Diploma level	8.0	8.0	8.2	7.9	7.7
	ISCED 5A, 6 – Bachelor degree and higher	62.5	63.4	63.0	62.5	63.6
Self White collar/Blue collar classification*	White collar high skilled	76.0	76.4	76.6	76.0	76.1
	White collar low skilled	12.5	12.2	12.0	12.5	12.6
	Blue collar high skilled	10.6	10.6	10.6	10.7	10.5
	Blue collar low skilled	0.9	0.8	0.8	0.9	0.8
Number of books at home*	0-10 books	4.9	4.4	4.0	4.9	5.0
	11-25 books	9.0	8.5	7.8	9.0	9.1
	26-100 books	28.6	28.4	27.5	28.6	28.6
	101-200 books	22.7	23.3	23.9	22.8	22.7
	201-500 books	21.1	21.4	21.8	21.0	21.0
	More than 500 books	13.7	14.0	15.0	13.6	13.6

Note: Variables marked with an * do not show missing values. Reported percentages exclude missing values, and so sum approximately to 100%, with rounding.

1 The percentage of Indigenous students in the full Pisa sample reported in LSAY Technical Report No. 43 *Sampling and Weighting of the 2003 LSAY Cohort* is 2.2%.
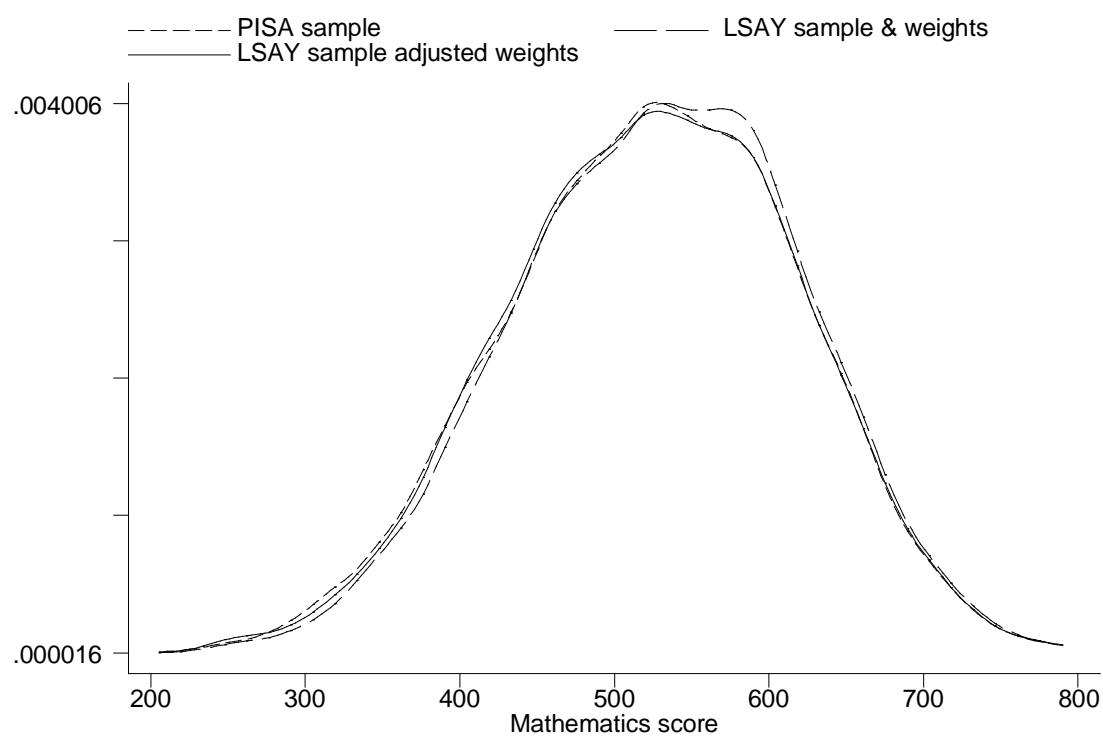
Source: Estimated from LSAY Y03 cohort and PISA 2003 data.

Table 2   Mean scores in the maths, literacy, scientific and problem solving domains on PISA for five plausible values with different weights

	Full Pisa sample	Pisa sample not in LSAY	Pisa sample in LSAY	LSAY weights	LSAY 2006 attrition weight	Revised LSAY 2003 student weight	Revised 2006 attrition weight
Mean plausible value in maths (pv1math)	524.3	497.6	530.1	529.1	534.4	524.9	525.3
Mean plausible value in reading	525.4	496.2	531.8	530.0	535.2	526.3	526.5
Mean plausible value in science	525.1	498.4	530.9	530.1	535.2	525.6	526.1
Mean plausible value in problem solving	529.8	503.6	535.6	534.5	539.6	530.4	530.8
Unweighted N	12,551	2,181	10,370	10,370	7,721	10,370	7,721

Source: Estimated from LSAY Y03 cohort and PISA 2003 data.

Figure 1   Distribution in 2006 on selected wave 1 variables of LSAY 2003 sample with PISA weights, LSAY weights, and revised weights

References

Jones, FL & McMillan, J 2001, 'Scoring Occupational Categories for Social Research: A Review of Current Practice, with Australian Examples', *Work, Employment and Society*, 15(3), 539—63.

Marks, GN & Long, M 2000, *Weighting the 1995 Year 9 cohort sample for differential response rates and sample attrition*, Longitudinal Surveys of Australian Youth technical paper no.15, ACER, Melbourne.

Rothman, S 2007, *Sampling and Weighting of the 2003 LSAY Cohort*, Longitudinal Surveys of Australian Youth technical report no.43, ACER, Melbourne.

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