13 Technical appendix

Respondents and collection methods

The use of multiple respondents in LSAC provides a rich picture of children's lives and development in various contexts. Across the first six waves of the study, data were collected from:

- parents of the study child:
  - Parent 1 (P1) – defined as the parent who knows the most about the child (not necessarily a biological parent)
  - Parent 2 (P2), if there is one – defined as another person in the household with a parental relationship to the child or the partner of Parent 1 (not necessarily a biological parent)
  - A parent living elsewhere (PLE), if there is one – a parent who lives apart from Parent 1 but who has contact with the child (not necessarily a biological parent)
- the study child
- carers/teachers (depending on the child's age)
- interviewers.

In earlier waves of the study, the primary respondent was the child's Parent 1. In the majority of cases, this was the child's biological mother, but in a small number of families this was someone else who knew the most about the child. Since Wave 2, the K cohort children have answered age-appropriate interview questions and, from Wave 4, they have also answered a series of self-complete questions. The B cohort children answered a short set of interview questions in Wave 4 for the first time. As children grow older, they are progressively becoming the primary respondents of the study.

A variety of data collection methods are used in the study, including:

- face-to-face interviews:
  - on paper
  - by computer-assisted interview (CAI)
- self-complete questionnaires:
  - during interview (paper forms, computer-assisted self-interviews (CASI) and audio computer-assisted self-interviews (ACASI)
  - leave-behind paper forms
  - mail-out paper forms
  - internet-based forms
- physical measurement of the child, including height, weight, girth, body fat and blood pressure

- direct assessment of the child's vocabulary and cognition
- time use diaries
- computer-assisted telephone interviews (CATI)
- linkage to administrative or outcome data (e.g. Medicare, MySchool).

Sampling and survey design

The sampling unit for LSAC is the study child. The sampling frame for the study was the Medicare Australia (formerly Health Insurance Commission) enrolments database, which is the most comprehensive database of Australia's population, particularly of young children. In 2004, approximately 18,800 children (aged 0–1 or 4–5 years) were sampled from this database, using a two-stage clustered design. In the first stage, 311 postcodes were randomly selected (very remote postcodes were excluded due to the high cost of collecting data from these areas). In the second stage, children were randomly selected within each postcode, with the two cohorts being sampled from the same postcodes.

A process of stratification was used to ensure that the numbers of children selected were roughly proportionate to the total numbers of children within each state/territory, and within the capital city statistical districts and the rest of each state. The method of postcode selection took into account the number of children in the postcode; hence, all the potential participants in the study Australia-wide had an approximately equal chance of selection (about one in 25). See Soloff, Lawrence, and Johnstone (2005) for more information about the study design.

Response rates

The 18,800 families selected were then invited to participate in the study. Of these, 54% of families agreed to take part in the study (57% of B cohort families and 50% of K cohort families). About 35% of families declined to participate (33% of B cohort families and 38% of K cohort families), and 11% of families (10% of B cohort families and 12% of K cohort families) could not be contacted (e.g. because the address was out-of-date or only a post office box address was provided).

This resulted in a nationally representative sample of 5,107 children aged 0–1 and 4,983 children aged 4–5, who were Australian citizens or permanent residents. This Wave 1 sample was then followed up at later waves of the study. Sample numbers and response rates for each of the main waves are presented in Table A.1.

Table A.1: Response rates, main waves, B and K cohorts, Waves 1–6

	Wave 1 (2004)	Wave 2 (2006)	Wave 3 (2008)	Wave 4 (2010)	Wave 5 (2012)	Wave 6 (2014)
B cohort						
Response rate of Wave 1	100.0%	90.2%	85.9%	83.0%	80.0%	73.7%
Response rate of available sample[a]	–	91.3%	88.2%	86.1%	83.5%	83.9%
Total (*n*)	5,107	4,606	4,386	4,242	4,077[b]	3,764
K cohort						
Response rate of Wave 1	100%	89.6%	86.9%	83.6%	79.4%	71.0%
Response rate of available sample[a]	–	90.8%	89.7%	87.2%	83.5%	80.5%
Total (*n*)	4,983	4,464	4,332[c]	4,164	3,952[c]	3,537
Total (B and K cohorts)						
Response rate of Wave 1	100%	89.9%	86.4%	83.3%	79.7%	72.3%
Response rate of available sample[a]	–	91.1%	89.0%	86.6%	83.5%	82.2%
Total (*n*)	10,090	9,070	8,718	8,406	8,029	7,301

Notes: This table refers to the numbers of parents who responded at each wave. [a] The available sample excludes those families who opted out of the study between waves. [b] B cohort: different numbers of parents and their children responded at Wave 5 (there were eight cases where a child interview was completed and the main interview with the parents was not). [c] K cohort: different numbers of parents and their children responded at Wave 3 (in one case a parent interview was completed and the interview with the study child was not); Wave 4 (in five cases a child interview was completed and the main interview with the parents was not); and Wave 5 (in four cases a child interview was completed and the main interview with the parents was not).

Sample weights

Sample weights (for the study children) have been produced for the study dataset in order to reduce the effect of bias in sample selection and participant non-response (Cusack & Defina, 2014; Daraganova & Sipthorp, 2011; Misson & Sipthorp, 2007; Norton and Monahan, 2015; Sipthorp & Misson, 2009; Soloff, Lawrence, & Johnstone, 2005; Soloff, Lawrence, Misson, & Johnstone, 2006). When these weights are used in the analysis, greater weight is given to population groups that are under-represented in the sample, and less weight to groups that are over-represented in the sample. Weighting therefore ensures that the study sample more accurately represents the sampled population.

These sample weights have been used in analyses presented throughout this report. Cross-sectional or longitudinal weights have been used when examining data from more than one wave. Analyses have also been conducted using Stata® *svy* (survey) commands, which take into account the clusters and strata used in the study design when producing measures of the reliability of estimates.

Overview of statistical methods and terms used in the report

Balanced panel

A balanced panel restricts the sample to individuals who have responded to the survey in all waves of the period under study. For example, a balanced panel for Waves 1–6 of the LSAC data consists of individuals who have responded in all six waves.

Confidence interval

A confidence interval (CI) is a range of values, above and below a finding, in which the actual value is likely to fall. The CI represents the accuracy of an estimate, and it can take any number of probabilities, with the most common being 95% or 99%. The analysis in this report uses a 95% confidence level. This means that the confidence interval covers the true value for 95 out of 100 of the outcomes.

Deciles, quartiles and quintiles

A decile is any of the nine values that divide data that have been sorted from lowest to highest into 10 equal parts, so that each part represents one-tenth of the sample or population. For example, the first decile of the income distribution cuts off the lowest 10% of incomes, and people in the first (or bottom) decile have the lowest 10% of incomes.

A quintile is any of the four values that divide data that have been sorted from lowest to highest into five equal parts. For example, people in the first (or lowest) income quintile have the lowest 20% of incomes.

A quartile is any of the three values that divide data that have been sorted from lowest to highest into four equal parts. For example, people in the first (or lowest) income quartile have the lowest 25% of incomes.

Mean

'Mean' is the statistical term used for what is more commonly known as the average – the sum of the values of a data series divided by the number of data points.

Odds ratios

An odds ratio (OR) is a measure of association between an exposure and an outcome. The odds ratio represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

ORs are used to compare the relative odds of the occurrence of the outcome of interest (e.g. disease or disorder), given exposure to the variable of interest (e.g. health characteristic, aspect of medical history). The OR can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

- OR = 1 Exposure does not affect odds of outcome
- OR > 1 Exposure associated with higher odds of outcome
- OR < 1 Exposure associated with lower odds of outcome.

Regression models

In statistical analysis, a regression model is used to identify associations between a 'dependent' variable (such as earnings) and one or more 'independent' or 'explanatory' variables (such as measures of educational attainment and work experience). In particular, it shows how the typical value of the dependent variable changes when any one of the independent variables is varied and all other independent variables are held fixed. Most commonly, regression models estimate how the mean value of the dependent variable depends on the explanatory variables – for example, mean (or 'expected') earnings given a particular level of education and work experience. Different types of regression models are used depending on factors such as the nature of the variables and data, and the 'purpose' of the regression model. The following types of models are estimated in this report:

Ordinary Least Squares models

Ordinary Least Squares models estimate linear associations between a dependent (or outcome) variable (such as earnings) and one or more independent (or explanatory) variables (such as age and educational attainment). The method finds the linear combination of the explanatory variables that minimises the sum of the squared distances between the observed values of the dependent variable and the values predicted by the regression model.

Logistic regression models

Logistic regression models are used to estimate the effects of factors, such as age and educational attainment, on a 'qualitative' or categorical dependent variable, such as labour force status, which is qualitative because it is not naturally 'quantitative' (or numerical), as is the case with income. The standard models examine 'binary' dependent variables, which are variables with only two distinct values, and estimates obtained from these models are interpreted as the effects on the probability the variable takes one of those values. For example, a model might be estimated on the probability an individual is employed (as opposed to not employed).

Statistical significance

In the context of statistical analysis of survey data, a finding is statistically significant if it is unlikely that the relationship between two or more variables is caused by something other than chance. That is, a relationship can be considered to be statistically significant if we can reject the 'null hypothesis' that hypothesizes that there is no relationship between measured variables. A common standard is to regard a difference between two estimates as statistically significant if the probability that they are different is at least 95%. However, 90% and 99% standards are also commonly used. The 95% standard is adopted for results presented in this report. Note that a statistically significant difference does not mean the difference is necessarily large, it simply means that you can be fairly confident that there is a difference.

References

Cusack, B., & Defina, R. (2014). *Wave 5 weighting & non-response* (Technical Paper No. 10). Melbourne: Australian Institute of Family Studies.

Daraganova, G., & Sipthorp, M. (2011). *Wave 4 weights* (Technical Paper No. 9). Melbourne: Australian Institute of Family Studies.

Misson, S., & Sipthorp, M. (2007). *Wave 2 weighting and non-response* (Technical Paper No. 5). Melbourne: Australian Institute of Family Studies.

Norton, A., & Monahan, K. (2015). *Wave 6 weighting and non-response (*Technical Paper No. 15). Melbourne: Australian Institute of Family Studies.

Sipthorp, M., & Misson, S. (2009). *Wave 3 weighting and non-response* (Technical Paper No. 6). Melbourne: Australian Institute of Family Studies.

Soloff, C., Lawrence, D., & Johnstone, R. (2005). *LSAC sample design* (Technical Paper No. 1). Melbourne: Australian Institute of Family Studies.

Soloff, C., Lawrence, D., Misson, S., & Johnstone, R. (2006). *Wave 1 weighting and non-response*. Melbourne: Australian Institute of Family Studies.