The Longitudinal Study of Australian Children: An Australian Government initiative

Data user guide

November 2015

Australian Government
Australian Institute of Family Studies

Contents

Abbreviations

ABS – Australian Bureau of Statistics

ACARA – Australian Curriculum, Assessment and Reporting Authority

ACASI - Audio Computer Assisted Self Interview

ACIR – Australian Childhood Immunisation Register

AEDC – Australian Early Development Census

AIFS – Australian Institute of Family Studies

ANU4 – Australian Nation University ranking of occupational prestige, 4th edition

ASCL – Australian Standard Classification of Languages

ANZSCO – Australian and New Zealand Standard Classification of Occupations

ASGC - Australian Standard Geographic Classification

ATSI – Aboriginal and Torres Strait Islander

BMI – Body Mass Index

BP – Study Child Blood Pressure

CA – Carer Allowance

CAI – Computer Assisted Interview

CAPI – Computer Assisted Personal Interview

CASI - Computer Assisted Self Interview

CATI – Computer Assisted Telephone Interview

CBC – Centre-Based Carer

CCB – Childcare Benefit

CSR – Child Self Report

DFRDB - Defence Forces Retirement and Death Benefits Scheme

DSP – Disability Support Pension

DSS – Department of Social Services

DVA – Australian Government Department of Veterans' Affairs

F2F – Parent 1 Face-to-Face Interview

FCF – Family Contact Form

FDC – Family Day Care

FDCQA – Family Day Care Quality Assurance

FTB – Family Tax Benefit

FTBA – Family Tax Benefit A

FTBB – Family Tax Benefit B

GPS – Global Positioning System

HBC – Home-Based Carer

IOBS – Interviewer Observations

IVF – In-Vitro Fertilisation

LDC – Long Day Care

LOTE – Language Other Than English

LSAC – Longitudinal Study of Australian Children

MBS – Medicare Benefit Scheme

MSN – Medicare Safety Net

MR – Matrix Reasoning test

NAPLAN – National Assessment Program—Literacy and Numeracy

NCAC – National Childcare Accreditation Council

NILF – Not In the Labour Force

NSA – Newstart Allowance

OMR – Optical Mark Recognition

OSHCQA – Outside School Hours Care Quality Accreditation

P1D – Parent 1 During Interview Questionnaire

P1L – Parent 1 Leave-Behind Questionnaire

P1SC – Parent 1 Self-Complete Questionnaire

P2L – Parent 2 Self-Complete Questionnaire

PBS – Pharmaceutical Benefit Scheme

PLE – Parent Living Elsewhere

PM – Physical Measurements

PPVT – Peabody Picture Vocabulary Test

PPVT-III – Peabody Picture Vocabulary Test, 3rd Edition

QIAS – Quality Improvement and Accreditation System (for Long Day Care centres)

ROC - Receiver Operating Characteristic

RSE – Relative Standard Error

SACC – Standard Australian Classification of Countries

SEIFA – Socio-Economic Indexes For Areas

SLI – Specific Language Improvement

SRS – Simple Random Sample

TIS – Telephone Interpreter Service

TQ – Teacher Questionnaire

TUD – Time Use Diary

WAI – Who Am I?

WISC – Wechsler Intelligence Scale for Children

Acknowledgements and citation

The current version of the LSAC Data User Guide has been updated by AIFS.

The Wave 6 data files were prepared by the ABS and AIFS.

Readers wishing to refer to this document should cite the following:

Australian Institute of Family Studies. (2015). *Longitudinal Study of Australian Children Data User Guide – November 2015*. Melbourne: Australian Institute of Family Studies.

1 Introduction

This data user guide is a reference tool for the users of the *Growing Up in Australia*: the Longitudinal Study of Australian Children (LSAC) datasets.

This document is intended to provide the necessary information to be able to use the LSAC data, and includes the survey methodology, file structure and variable naming conventions. Particular issues are highlighted to ensure data analysts apply the LSAC data appropriately in their research.

The following documentation is also useful to data users and is available on the study website <www.growingupinaustralia.gov.au/index.html>:

- questionnaires and interview specifications marked with variable names (including mock questionnaires for Computer Assisted Interview (CAI) instruments)[1]
- a data dictionary
- technical papers on weighting, non-response and other issues
- data issues papers.

Data users should read the "Important issues for data analysis" section carefully. It outlines particular aspects of the sample design that have important implications for interpreting analyses from the study.

Further information on data usage is contained in the *Data users information* pages on the LSAC website <www.growingupinaustralia.gov.au/data/>.

Feedback about this data user guide is welcome. If there is something that you would find useful that is currently not included, or if you had difficulty understanding any of the guide's content, please let us know by emailing us at: aifs-lsac@aifs.gov.au.

[1] Feedback from data users suggests that marked questionnaires with interview specifications are often the best way to find sections relevant to proposed research topics, and to illustrate the breadth of information available in the study.

2 What is LSAC?

Growing Up in Australia: the Longitudinal Study of Australian Children (LSAC) continues to examine the impact of Australia's unique social and cultural environment on the next generation. The study aims to build upon understandings of child development, inform social policy debate and identify opportunities for intervention and prevention strategies in policy areas concerning children and their families.

2.1 Objectives

LSAC explores family and social issues while addressing a range of research questions about children's development and wellbeing. Information is collected on children's physical and mental health, education and social, cognitive and emotional development. The data is sourced from parents, child carers, pre-school and school teachers and the children themselves.

The study's longitudinal structure enables researchers to determine critical periods for providing services and welfare support, and to identify long-term consequences of policy innovations (for more details see LSAC Discussion Paper No.1, *Introducing the Longitudinal Study of Australian Children*).

LSAC provides a database to help develop a comprehensive understanding of children's development in the context of Australia's social, economic and cultural environment. The study is the first ever comprehensive, national Australian data collection on children as they grow up.

2.2 Who is involved?

LSAC is undertaken in partnership between the Department of Social Services (DSS), the Australian Institute of Family Studies (AIFS) and the Australian Bureau of Statistics (ABS), with advice provided by a consortium of leading researchers known as the LSAC Consortium Advisory Group.

The Wave 1 data collection was undertaken for AIFS by private social research companies Colmar-Brunton Social Research and I-view/NCS Pearson. Data collection for Waves 2-6 was undertaken by ABS.

2.3 Timelines

Development for the study commenced in March 2002 with a testing phase involving over 500 families that continued through 2003. Recruitment for the main study took place between March and November 2004, and over 10,000 children and their families agreed to participate. From 2004, participating families have been interviewed every two years, and between-wave mail-out questionnaires were sent to families in 2005 (Wave 1.5), 2007 (Wave 2.5) and 2009 (Wave 3.5). Additional between-wave questionnaires (Waves 4.5 and 5.5) were undertaken via online web forms from 2009 for the purposes of updating the contact details of study participants.

2.4 Sample design

The focus of the study is on the developmental pathways of Australian children, so the study child is the sampling unit of interest. A dual cohort cross-sequential design was adopted as shown in Figure 1.

Cohor	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 4**	**Wave 5**	**Wave 6**
B cohort	0–1 yrs	2–3 yrs	4–5 yrs	6–7 yrs	8–9 yrs	10–11 yrs
K cohort	4–5 yrs	6–7 yrs	8–9 yrs	10–11 yrs	12–13 yrs	14–15 yrs

Figure 1: The dual cohort cross-sequential design of LSAC

Two cohorts of children were selected from children born within two 12-month periods:

- B cohort (infant cohort): children born March 2003–February 2004
- K cohort (child cohort): children born March 1999–February 2000

Further information about the design of the sample is available in the "Survey methodology" section of this guide, and in LSAC Technical Paper No. 1, "*Sample Design*" (available from the study website, <www.growingupinaustralia.gov.au/pubs/technical/index.html>).

2.5 Study informants

The study collects data from multiple informants:

- Study child is the cohort child, themselves.
- Parent 1 (P1) is defined as the parent who knows the study child best; in most cases this is the child's biological mother.
- Parent 2 (P2) is Parent 1's partner or another adult in the home with a parental relationship to the study child; in most cases this is the biological father, but step-fathers are also common.
- Parent living elsewhere (PLE) is a parent who does not live with the study child; most commonly the biological father after separating from the biological mother. This collection was started in Wave 2.
- Teachers and childcare workers involved with the study child.

In addition, data are linked to the file from the National Childcare Accreditation Council, Medicare Australia, ABS and the National Assessment Program—Literacy and Numeracy (NAPLAN).

2.6 Mother/Father data

While Parent 1 is usually the mother and Parent 2 is usually the father, this is not always the case. However, many users prefer to analyse the data by parent gender (i.e., Mother and Father rather than Parent 1 and Parent 2). Therefore, all the variables collected for both Parent 1 and Parent 2 are also presented as Mother and Father variables. Note that Parent 1 and Parent 2 may be the guardians of the child and not the child's biological parents. In this context, Mother should be taken to mean "female parent/guardian". Sometimes Parent 1 (and/or Parent 2) might change between waves. For instance, Parent 1 may be reported as female across subsequent waves, although the parent may, in fact, be different people.

If there are two female parents, Parent 1 is coded as Mother and Parent 2 is coded as Father. This will be maintained if the parents swap between Parent 1 and Parent 2 in subsequent waves. This means that there are a small number of female Fathers that analysts should be mindful of when working with these variables.

3 Instruments

Table 1 summarises the data collection instruments used in each wave.

Table 1: Data collection modes by wave

Questionnaire	Mode	Completed by	Indicator variable	W1	W2	W3	W4	W5	W6
Face-to-face interview (F2F)	Paper	Parent 1	N/A	BK	-	-	-	-	-
Face-to-face interview (F2F)	Computer	Parent 1	N/A	-	BK	BK	BK	BK	BK
Parent 1 during interview (P1D)	Paper	Parent 1	[*]p1dd	BK	BK	BK	-	-	-
Parent 1 during interview (CASI)	Computer	Parent 1	[*]p1dd	-	-	-	BK	BK	BK
Parent 1 leave behind (P1L)	Paper	Parent 1	[*]p1scd	BK	BK	BK	-	-	-
Parent 2 leave behind (P2L)	Paper	Parent 2	[*]p2scd	BK	BK	BK	BK	BK	BK
Child self-report (CSR)	Computer	Study Child	[*]csrd	-	K	K	B	BK	BK
Audio computer-assisted self-interview (ACASI)	Computer	Study Child	Need consent from: P1 [*]id40e & SC [*]id40f	-	-	-	K	K	K
Time use diary (TUD)	Paper	Parent 1	N/A	BK	BK	BK	-	-	-
Time use diary (TUD)	Computer	Study Child	Need consent from: P1 [*]id40i & SC [*]id40j	-	-	-	K	K	K
Parent living elsewhere (PLE)	Paper—mailed out	PLE	[*]plescd	-	BK	-	-	-	-
Parent living elsewhere (PLE CATI)	Computer/ Telephone	PLE	[*]plescd	-	-	BK	BK	BK	BK
Home-based carer (HBC)	Paper	Carer	[*]hbccbc	B	B	-	-	-	-
Centre-based carer (CBC)	Paper	Carer	[*]hbccbc	B	B	-	-	-	-
Teacher questionnaire (TQ)	Paper	Teacher	[*]tcd	K	K	BK	BK	BK	BK
Physical measurements (PM)	Computer	Study Child	Need consent from: P1 [*]id30d & SC [*]id30e	BK	BK	BK	BK	BK	BK
Who am I? (WAI)	Computer	Study Child	cid44a1	K	-	B	-	-	-
PPVT assessment (PPVT)	Computer	Study Child	[*]ppvtd	K	K	BK	B	B	-
Matrix reasoning (MR)	Computer	Study Child	[*]id44a1	-	K	K	BK	B	B
Study child blood pressure (BP)	Computer	Study Child	Need consent from: P1 [*]id47a & SC [*]id47b	-	-	-	K	K	B
Interviewer observations (IOBS)	Computer	Interviewer		BK	BK	BK	BK	BK	BK
Executive functioning (EXEC - CogSTATE)	Computer	Study Child	[*]id40m [*]id40n	-	-	-	-	-	K

Notes: The indicator variable can be used to see if data is present or not for a particular instrument (for more information see sections 8.6 & 8.7). The [*] in the indicator variable should be replaced by the age indicator (a, c, d, e, f, g or h) as discussed below. In-between waves were administered using mail out surveys for Waves 1.5, 2.5 and 3.5. Waves 4.5 and 5.5 used online web forms to update contact details.

The following methods are used to collect study data.

- The face-to-face interview (F2F) is conducted with Parent 1 (although, in Wave 1, Parent 2 could complete some sections if this was more convenient). This component is undertaken with all participating families at each wave. Some interviews might be completed over the telephone in full; for example, with participating families in remote areas (see section 11.3.7).
- The Parent 1 during interview questionnaire (P1D) consists of self-complete items for which it was considered important to achieve high response rates. In Wave 4 it became a computer-assisted self-interview (CASI).
- The Parent 1 leave-behind questionnaire (P1L) consists of lower priority self-complete items. Efforts are made to obtain this data from Parent 1 while the interviewer is in the home. This form became part of the CASI.
- The Parent 2 leave-behind questionnaire (P2L) consists of self-complete items. Efforts are made to obtain this data from Parent 2 while the interviewer is in the home. If this is not possible the questionnaire is left for completion at a later time.
- Child self-report interview (CSR) consists of items answered by the study child. For children younger than 10 years old it is administered by an interviewer. For children 10 years and older it is administered via audio computer-assisted self-interview (ACASI). As part of the interview, physical measurements are taken and other assessments (such as measures of cognition or achievement) are administered to the study child.
- The study child completes an audio computer-assisted self-interview (ACASI) by themselves. This method allows sensitive content to be answered by the child in total anonymity.
- The time use diary (TUD) documents a 24-hour period of the child's life. In Waves 1, 2 and 3, the child's family were asked to complete two TUDs, one for a weekday and one for a weekend day. A different procedure was implemented in Wave 4. In Wave 4, the study child (K cohort only) was asked to complete one TUD. A TUD form with instructions on how and when to fill it in was sent to the study child prior to the interview. The study child was asked to fill in the TUD form on the day before the interview date. The next day, during the interview, the interviewer asked the child to describe "yesterday" using the TUD form. The day the diary referred to could be any day of the week depending on when the interview was scheduled.
- The parent living elsewhere questionnaire (PLE) was first included in Wave 2 as a mail-back questionnaire. In Wave 3 it became a computer-assisted telephone interview (CATI).
- The home-based carer questionnaire (HBC) is for children aged 0-1 and 2-3 years who receive childcare in a home environment, most commonly from a grandparent.
- The centre-based carer questionnaire (CBC) is for children aged 0-1 and 2-3 years who receive childcare from long day care programs in centres, schools, occasional care programs, multi-purpose centres and other arrangements.
- The teacher questionnaire (TQ) is for children aged 4-5 years and older who attend a school or, for some 4-5 year olds, a preschool or long day care centre.
- Interviewers make observations (IOBS) with permission of the respondent about the interview, state of the house (where the interview was conducted) and the neighbourhood characteristics of where the respondent lives.
- In Wave 1 the Australian Early Development Census (AEDC) was included as a nested study, which involved the AEDC questionnaire being sent with the LSAC K cohort teacher questionnaire in Victoria, Queensland and Western Australia. The AEDC is a community-level measure of young children's development based on a teacher-completed checklist. It consists of over 100 questions measuring five developmental domains: language and cognitive skills; emotional maturity; physical health and wellbeing; communication skills and general knowledge; and social competence. More information can be found on the AEDC website: <www.aedc.gov.au>.
- The family contact form (FCF) recorded information about any contact between the interviewer and the family of each of the selected children at the time of Wave 1, regardless of whether they agreed to participate in the study or not. The information was mainly used by the fieldwork agency, with the only information from the FCF available in the publicly released dataset being the information on the family's home and neighbourhood. In subsequent waves, this information was included as part of the interviewer observations of the face-to-face interview.

- Between-wave questionnaires (Wave 1.5, Wave 2.5 and Wave 3.5) are brief questionnaires sent to respondents to complete and return in the year between main waves of data collection. Between-wave surveys help to maintain contact with study families and collect information about activities and development in the year between the main waves. For Waves 4.5 and 5.5, online web forms were used to update contact details of study participants.

3.1 Child assessments

3.1.1 Physical measurements

Weight

For the B cohort in Wave 1, the child's weight was obtained by calculating the difference between the weight of Parent 1 (or another adult) with the child and the weight of the parent/other adult on their own. For the B cohort at all subsequent waves, and the K cohort at all waves, the child's weight was measured directly.

In Wave 1 the scales used were Salter Australia glass bathroom scales (150 kg x 50 g). In Waves 2 and 3, these scales were used along with HoMedics digital BMI bathroom scales (180 kg x 100 g). In Waves 4, 5 and 6, Tanita body fat scales were used.

Height

Height is measured for children aged 2 years and older. In Waves 1, 2 and 3, height was measured using an Invicta stadiometer, from Modern Teaching Aids. In Waves 4, 5 and 6, a laser stadiometer was used. Two measurements were taken, and if the two measurements differed by 0.5 cm or more, a third measurement was taken. The average of the two closest measures was included on the data file.

Girth

This measurement is taken for children aged 2 years and older using a non-stretch dressmaker's tape, positioning the tape horizontally over the navel. In all waves, two measurements were taken, and if these differed by 0.5 cm or more, a third measurement was taken. The average of the two closest measures was recorded on the data file.

Body fat

A body fat measurement was included in Waves 4, 5 and 6, with the reading provided by the same scales used for weight (Tanita body fat scales). Issues with the body-fat measurement are outlined in the Data Issues Paper.

Head circumference

This measurement was only taken for the B cohort in Wave 1, using an Abbott head circumference tape. Two measurements were taken, and if these differed by 0.5 cm or more, a third measurement was taken. The average of the two closest measures was included on the data file.

Blood pressure

This measurement was taken for the K cohort in Waves 4 and 5 and for the B cohort in Wave 6 using the A&D Digital Blood Pressure Monitor - Model UA-767. The interviewer took two measurements, with a one-minute interval between the measurements. Both of the readings were included in the data file.

3.1.2 "Who am I?" (WAI)[2]

The "Who am I?" assessment is a direct child assessment measure that requires children to copy shapes (a circle, triangle, cross, square and diamond) and write numbers, letters, words and sentences. For the LSAC testing, there was a change to "Who Am I?", Item 11: "This is a picture of me" was replaced with a sentence to be copied, "John

[2] The "Who Am I?" is copyrighted by the Australian Council for Educational Research, Melbourne, 1999.

is big". The "Who am I?" assessment was used for children aged 4-5 years (Wave 1 K cohorts and Wave 3 B cohorts) to assess the general cognitive abilities needed for beginning school.

The study child was given his/her own answer booklet to draw and write in. What they wrote/drew was assessed by experienced researchers at the Australian Council for Educational Research (ACER). See *Data Issues Waves 1 to 6* for details of the Rasch Modelling used to score the WAI (available on the study website <www.growingupinaustralia.gov.au/pubs/issues/index.html>.

3.1.3 Peabody Picture Vocabulary Test (PPVT)[3]

A short form of the Peabody Picture Vocabulary Test (PPVT-III), a test designed to measure a child's knowledge of the meaning of spoken words and his or her receptive vocabulary for Standard American English, was developed for use in the study. This adaptation is based on work done in the USA for the Head Start Impact Study, with a number of changes made for use in Australia.

Different versions of the PPVT containing different, although overlapping, sets of items of appropriate difficulty were used for the children at ages 4-5, 6-7 and 8-9 years. A book with 40 plates of display pictures was used. The child points to (or says the number of) a picture that best represents the meaning of the word read out by the interviewer.

Scores are created via Rasch Modelling so that changes in scores represent real changes in functioning, rather than just changes in position relative to peers. See *Data Issues Waves 1 to 6* for more details (available from <www.growingupinaustralia.gov.au/pubs/issues/index.html>.

3.1.4 Matrix Reasoning[4]

Children completed the Matrix Reasoning (MR) test from the Wechsler Intelligence Scale for Children, 4th edition (WISC-IV) at ages 6-7, 8-9 and 10-11 years. This test of non-verbal intelligence presents the child with an incomplete set of diagrams (an item) and requires them to select the picture that completes the set from five different options. The data file includes raw scores (number of correct responses) and scaled scores based on age norms given in the WISC-IV manual. The instrument comprises 35 items of increasing complexity. Children start on the item corresponding to their age-appropriate start point. If a child does not answer correctly on the first or second start-point items, the examiner should ask two items prior to the age-appropriate start point (called "reverse administration"). Reverse administration was not implemented in the LSAC instrument. See the discussion of this issue in See *Data Issues Waves 1 to 6* (available from <www.growingupinaustralia.gov.au/pubs/issues/index.html>).

3.1.5 Executive functioning (EXEC/CogState)

The executive functioning of children in the K cohort was tested from Wave 6 using the Groton Maze Learning Test (GML) (See <cogstate.com/computerized-tests/cognitive-tasks/groton-maze-learning-test/>). The outcome variables are contained in the CogState dataset, where a series of cognitive testing batteries have been customised for use in LSAC. Each row of a CogState dataset represents one task in the CogState test battery for one study subject in one test session. Each column represents demographic information or an outcome variable.

The GML test contains five learning trials (i.e. the subject repeats the same task five times), where the child is shown a 10 x 10 grid of tiles on a computer touch-screen. A 28-step pathway is hidden among these 100 possible locations. The child is instructed to move one step from the start location and then to continue, one tile at a time, toward the end. The subject repeats the task while trying to remember the pathway they have just completed and learns the 28-step pathway though the maze on the basis of trial and error feedback. The scores are interpreted by

---

[3] The Peabody Picture Vocabulary Test, Third Edition (PPVT-III) Form IIA is copyrighted by Lloyd Dunn, Leota Dunn, Douglas Dunn, American Guidance Service, Inc., 1997, and published exclusively by AGS Publishing. Permission to adapt and create a short form for LSAC was granted by the publisher. The PPVT - III - LSAC Australian Short-form was developed by S. Rothman, Australian Council for Educational Research (ACER), Melbourne, from the Peabody Picture Vocabulary Test, Third Edition (PPVT-III), Form IIA, English edition.

[4] The Wechsler Intelligence Scale for Children, Fourth Edition is copyrighted by Harcourt Assessment, Inc., 2004.

calculating the total number of errors made in attempting to learn the same hidden pathway. A lower score indicates better performance.

3.1.6 Rice Test of Grammaticality Judgement (GJT/SLI)

As children grow older different methods are needed to assess the presence or absence of specific language impairment (SLI). That is, to identify whether children are meeting expected performance levels in achieving the adult standard of English grammar. Where LSAC children were identified in early waves to have poor language performance, it was not possible to distinguish the children with and without SLI. The Rice Grammaticality Judgement Task (GJ Task) was therefore introduced in Wave 6 for children of the K cohort.

The GJ Task is a short, automated (administered by ACASI) task that requires the study child to distinguish between grammatical and non-grammatical utterances known to be vulnerable to SLI in English-speaking children (Rice, Hoffman and Wexler, 2009). The study child listens through earphones as 20 pre-recorded items are spoken and enters their response by clicking the appropriate radio buttons (*1* for "Right", *5* for "Not so good", and *9* for "Hear again"). Its sensitivity and specificity for SLI are .70 with a ROC of approximately 0.85.

3.2 Response rates

The number and percentages of survey instruments of each type that were completed at each wave are shown in Table 2. More detailed information on non-response can be found in the technical papers on weighting and non-response (available at <www.growingupinaustralia.gov.au/pubs/technical/index.html>.

Table 2: Waves 1-6 instrument response

	B cohort			K cohort		
Wave 1 instrument[a]	Eligible[b]	Actual[c]	%	Eligible[b]	Actual[c]	%
F2F	5,107	5,107	100	4,983	4,983	100
P1L	5,107	4,341	85	4,983	4,229	85
P2L	4,630	3,696	80	4,286	3,388	79
TUD 1	5,107	4,031	79	4,983	3,867	78
TUD 2	5,107	3,751	73	4,983	3,582	72
WAI	N/A	N/A	N/A	4,983	4,880	98
PPVT	N/A	N/A	N/A	4,983	4,382	88
HBC	788	342	43	N/A	N/A	N/A
CBC	436	233	53	N/A	N/A	N/A
TQ	N/A	N/A	N/A	4,761	3,276	69
AEDC	N/A	N/A	N/A	1,366	720	53
W1.5	5,061	3,573	71	4,935	3,594	73

Wave 2 instrument[a]	B cohort			K cohort		
	Eligible[b]	Actual[c]	%	Eligible[b]	Actual[c]	%
F2F[d]	5,107	4,606	90	4,983	4,464	90
P1D	4,606	4,504	98	4,464	4,358	98
P1L	4,606	3,536	77	4,464	3,495	78
P2L	4,099	3,128	76	3,804	2,949	78
TUD 1	4,606	3,477	75	4,464	3,446	77
TUD 2	4,606	3,459	75	4,464	3,460	78
PPVT	N/A	N/A	N/A	4,464	4,409	99
MR	N/A	N/A	N/A	4,464	4,402	99
PLE mail-out	400	96	24	612	199	33
HBC	791	533	67	N/A	N/A	N/A
CBC	1672	1,144	68	N/A	N/A	N/A
TQ	N/A	N/A	N/A	4,447	3,632	82
W2.5	5,107	3,246	64	4,983	3,252	65

Wave 3 instrument[a]	B cohort			K cohort		
	Eligible[b]	Actual[c]	%	Eligible[b]	Actual[c]	%
F2F[d]	5,107	4,386	86	4,983	4,331	87
P1D	4,386	3,831	87	4,331	3,807	88
P2L	3,900	2,753	71	3,707	2,680	72
TUD 1	4,386	2,959	67	4,331	2,961	68
TUD 2	4,386	2,950	67	4,331	2,963	68
PPVT	4,386	4,266	97	4,331	4,273	99
WAI	4,386	4197	96	N/A	N/A	N/A
MR	N/A	N/A	N/A	4,331	4,270	99
PLE CATI	346	272	77	510	403	79
TQ	4,114	3,395	83	4,275	3,643	85

Wave 4 instrument[a]	B cohort			K cohort		
	Eligible[b]	Actual[c]	%	Eligible[b]	Actual[c]	%
F2F[d]	5,107	4,242	82	4,983	4,164	84
CASI	4,242	4,210	99	4,164	4,116	99
P2L	3,706	2,677	72	3,512	2,645	75
CSR	4,242	4,181	99	N/A	N/A	N/A
ACASI	N/A	N/A	N/A	4,169*	4,094	99
TUD	N/A	N/A	N/A	4,169*	3,994	96
PPVT	4,242	4,185	99	N/A	N/A	N/A
MR	4,242	4,180	99	4,169*	4,103	99
PLE CATI	439	377	86	572	493	86
TQ	4,143	3,427	83	4,025	3,352	83

Wave 5 instrument[a]	B cohort			K cohort		
	Eligible[b]	Actual[c]	%	Eligible[b]	Actual[c]	%
F2F[d]	5,107	4,085	80	4,983	3,956	79
CASI	4,077	4,010	98	3,952	3,857	98
P2L	3,512	2,444	70	3,277	2,333	71
CSR	4,026*	4,014	100	3,872	3,850	99
ACASI	N/A	N/A	N/A	3,873*	3,844	99
TUD	N/A	N/A	N/A	3,871*	3,649	94
PPVT	4,026	3,977	99	N/A	N/A	N/A
MR	4,027	3,985	99	N/A	N/A	N/A
PLE CATI	537	404	75	614	464	76
TQ	4,021	3,490	87	3,857	3,225	84
Wave 6 instrument[a]	**B cohort**			**K cohort**		
	Eligible[b]	Actual[c]	%	Eligible[b]	Actual[c]	%
F2F[d]	5,107	3,764	74	4,983	3,537	71
CASI	3,759	3,668	98	3,526	3,376	96
P2L	3,197	2,311	72	2,904	2,212	76
CSR	N/A	N/A	N/A	3,388	3,317	98
ACASI	3,648*	3,597	99	3,386*	3,313	98
TUD	3,649*	3,460	95	3,387*	3,071	91
EXEC	N/A	N/A	N/A	3,386*	3,333	98
GJT	N/A	N/A	N/A	3,386*	3,281	97
MR	3,648*	3,585	98	N/A	N/A	N/A
PLE CATI	559	398	71	554	420	76
TQ	3,678	3,102	84	3,422	2,698	79

Notes: * Represents instances where a child interview was completed and the main interview with the parents was not. Specifically, in Wave 4 there were five cases (K cohort). In Wave 5 there were eight cases for the K cohort and four cases for the B cohort. In Wave 6 there were eleven cases for the K cohort and four cases for the B cohort. N/A = Not administered

[a] Questionnaire acronyms are detailed above in section 3, Table 1: Data collection modes by wave.

[b] "Eligible" means the number of LSAC children for whom a questionnaire was applicable (e.g., children are eligible for a HBC questionnaire if the child's main care is attended for 8 hours or more per week and this is home-based care).

[c] "Actual" means the number of respondents for whom a form was returned.

d Response rates for Waves 2 to 6 as a proportion of Wave 1 families.

3.2.1 Parent 1 questionnaires

In Wave 1, interviewers encouraged the parents to complete the P1L and P2L forms while the interviewer was in the home. Interviewers were also able to pick up forms in some cases, when forms were left behind. Forms not given to interviewers were mailed back. Two reminders were made for forms that were not returned.

In Wave 2, Parent 1 had two forms to complete. Interviewers were instructed that the P1D form *must* be completed when they were in the home (resulting in a high response rate). The P1L was generally left behind to be mailed back, as there was not enough time for these to be completed. Interviewers were generally not required to pick up the forms. Up to four reminders were made for forms that were not returned; however, the P1L forms showed lower response rates in Wave 2 compared with Wave 1. This may have been because P1 had already completed one form or because interviewers did not generally pick up forms.

For Wave 3, there was only one Parent 1 self-complete form. Interviewers were instructed that this form must be completed while the interviewer was in the home. However, only two thirds of parents were able to do so. Three reminders were given for forms not returned.

In Wave 4, Parent 1 was asked to complete a CASI, which resulted in a response rate of 99% of eligible respondents. This was higher than the response rate of 88% of eligible respondents achieved in Wave 3 using the self-complete form.

In Wave 5, response rates were very similar to response rates obtained in Wave 4. This was due to no mode changes and attrition tapering off.

In Wave 6, response rates are similar to previous waves using the same mode. There is a slight decrease from the K cohort completion of the CASI from 98% in Wave 5 to 96% in Wave 6.

3.2.2 Parent 2, TUD and teacher forms

Response rates to the P2L and the TUD were broadly similar between waves (Waves 1, 2 and 3) at between 67 and 79%, while the carer and teacher questionnaire response rates were much improved in Wave 2, with similar response rates at Wave 3. In Wave 4 the TUD response rate was 96%. The higher response rate could be contributed to changes in the procedure and in the informant. In Waves 4, 5 and 6, the interviewer collected the TUD information from the child instead of the parent. The data were collected as part of the interview rather than leaving a diary that previously required completion and return via mail by respondent families after the visit.

3.2.3 PLE response

The PLE questionnaire was introduced in Wave 2 and applies to children who see their "parent living elsewhere" (PLE) at least once a year. There are three stages at which non-response can occur: (1) obtaining contact details from Parent 1; (2) obtaining permission from Parent 1; and (3) receiving a response from the PLE.

In Wave 2, contact details were given for 69% of cases for the B cohort and 70% of cases for the K cohort, and responses were received from 35% of PLEs sent a questionnaire for the B cohort and 47% for the K cohort.

Due to the relatively low response in Wave 2 to the mail-out questionnaire, a change in methodology was introduced in Wave 3. Where Parent 1 had provided contact details, PLEs were telephoned and asked to respond to a computer-assisted telephone interview (CATI). The response from PLEs who were approached was very positive. Of the 856 PLEs that interviewers attempted to contact, interviews were achieved with 675 (79%) and only 53(6%) refused an interview. Most of the remaining non-responses were due to not being able to contact the PLE.

In Wave 3, Parent 1 was explicitly asked for their permission to contact the PLE. Therefore, it was easy for Parent 1 to refuse to provide any information about the PLE or refuse the PLE's participation. This meant that no information was obtained for 260 (18%) PLEs.

It is worth noting that from Wave 2 onwards, there was no direct question asking the Parent 1 permission to contact the PLE: some Parent 1 respondents refused the PLE's participation.

Table 3 summarises the situation with regard to PLEs for Waves 3 to 6.

Table 3: Waves 3 to 6: Information obtained with regard to PLE

	Wave 3			Wave 4			Wave 5			Wave 6		
	B	K	Total	B	K	Total	B	K	Total	B	K	Total
PLE identified during P1 interview	578	837	1415	674	878	1,552	773	911	1,684	778	817	1,595
Eligible PLE*	346	510	856	439	572	1,011	537	614	1,151	559	554	1,113

Note: *The PLE is considered eligible when: (1) the PLE satisfies the parental requirements; i.e. PLEs who see the study child at least once a year; (2) the PLE's contact details are available; (3) Parent 1 did not explicitly refuse permission to contact the PLE.

4 The LSAC data release

Data users are required to read this user guide for access to and use of DSS longitudinal survey datasets. Users are also required to complete a dataset application form and sign a deed of license. Users must abide by strict security and confidentiality protocols. Instructions on how to access data can be found on the LSAC website at <www.aifs.gov.au/growingup/data/index.html>.

4.1 Data security requirements

The deed of licence stipulates numerous security requirements for the data, including:

- The LSAC CD-ROM MUST be kept secure in a locked filing cabinet or other secure container when not in use.
- The LSAC data (and any derivatives of the LSAC data) MUST be stored on a password-protected computer or network.
- Your password MUST include a mixture of uppercase and lowercase characters, be at least eight characters long and include some non-alphanumeric characters such as #, ;, *, etc.
- Any printed unit record output MUST be stored in a locked filing cabinet or other secure container when not in use. Any printed unit record output MUST be shredded if no longer required.
- LASC data users MUST NOT provide the unit record data to any unauthorised individual.
- There MUST be a means of limiting access to the work area where the data are kept and there MUST be tamper evident barriers to access (i.e. if there were a break-in, it would be obvious from broken glass, a damaged lock, etc.).
- If a data user has an individual license and changes employers, they MUST inform DSS prior to doing so: LSAC data MUST NOT be left with a former employer unless authority has been granted by DSS to do so. Data may be able to move with the individual, depending on the research to be undertaken and the new employer.
- If a data user embarks on a new research project, permission MUST be sought from DSS to use the data for the new project.

4.2 How data files are provided

All data are provided in three formats—SAS, SPSS and STATA—however, users can transfer the data to other formats if they wish. The CD-ROM and/or website also include extensive data documentation including this document, marked-up questionnaires and variable frequencies. The data files and the other documentation are discussed in detail in later sections of this document.

5 File structure

For the Wave 6 data release, the following datasets are available:

- 12 datasets comprising the main datasets for each wave and cohort (lsacgrb0,[5] lsacgrb2, lsacgrb4, lsacgrb6, lsacgrb8, lascgrb10, lsacgrk4, lsacgrk6, lsacgrk8, lsacgrk10, lsacgrk12 and lsacgrk14)
- 21 time use diary datasets:
  - one cleaned datafile with problematic cases deleted for each cohort for Waves 1, 2 and 3 (diaryb0, diaryb2, etc.)
  - one datafile with the cases deleted from the above files after cleaning for each cohort for Waves 1, 2 and 3 (poortudsb0, poortudsb2, etc.)
  - one datafile with all cases and no data cleaning performed on them for each cohort for Waves 1, 2 and 3 (ucdiaryb0, ucdiaryb2, etc.)
  - one datafile for K cohort only for Wave 4 (tudk10)
  - one datafile for K cohort only for Wave 5 (tudk12)
  - one datafile for K cohort only for Wave 6 (tudk14)
- three Medicare Australia datasets representing information from the three Medicare Australia databases the information was drawn from (mbs, pbs and acir)
- two study child household composition datasets, one for each cohort (hhgrb, hhgrk)
- two PLE household composition datasets, one for each cohort (plehhgrb, plehhgrk)
- two Wave 2.5 datasets, one for each cohort (lsacgrb3, lsacgrk7)
- two Wave 3.5 datasets, one for each cohort (lsacgrb5, lsacgrk9)
- the LSAC NAPLAN dataset (lsacnaplan)
- the LSAC MySchool dataset (lsacmyschool)
- the AEDC dataset (aedc) NB this is available with additional approval.

Note: Wave 1.5 datasets have been added to the Wave 1 datasets. This is possible because all participants who responded at Wave 1.5 had to complete a Wave 1 interview. This is not the case with the other between-wave mail-outs, as respondents may have completed any prior combination of interviews.

This structure has been used to reduce the size of the main datasets and because some data are formatted using more than one record for each child.

5.1 Main dataset

The main dataset consists of the data from all questionnaires except the time use diary, Wave 2.5, Wave 3.5, Wave 4.5, Wave 5.5, some household composition information and LSAC NAPLAN data. Data from the instruments are presented in the following order:

- FCF (Wave 1 files only)
- F2F
- P1 self-complete (except Wave 1 files)
- P2 self-complete
- PLE self-complete/interview (except Wave 1 files)
- Teacher/Carer questionnaire[6]

---

[5] File names in this section are for the general release datasets (see section 9: Confidentialisation below), users of the in-confidence data should substitute “ic” for “gr” in the file names.

[6] Since the CBC or HBC forms were only dispatched to the child’s main care type, each child could only have one of these completed for them. Hence, for Waves 1 and 2, HBC and CBC data are merged into a single set of variables where possible. This data is given in the order of the HBC questionnaire, with questions appearing only in the CBC form given at the end.

- Wave 1.5 data (Wave 1 files only)

A number of derived variables are included in the output dataset alongside the raw responses used in their derivation. Additionally the main datasets contain status variables (e.g., date of interview, whether each type of form was returned, etc.), ABS Population Census and NCAC data, and weights.

5.1.1 Australian Bureau of Statistics Census of Population and Housing data

Public data from the Australian Bureau of Statistics Census of Population and Housing have been added to the file to enhance the range of neighbourhood characteristics available for analysis with the LSAC data. Census data is available for the child's residence from Waves 1 to 6.

The items currently included are:

- SEIFA – rounded off to the nearest 10 for on the general release file
- Remoteness Area classification
- percentage of persons aged under 5, 10 and 18 years
- percentage of persons born in Australia
- percentage of persons speaking English-only at home
- percentage of persons with Aboriginal and Torres Strait Islander (ATSI) origins
- percentage of persons who completed Year 12 schooling
- percentage of persons in above-median income category
- percentage of persons working
- percentage of households with Internet capacity (in 2006 Census only)
- Percentage of households with broadband (in 2006 Census only).

Census data is either linked at the Statistical Local Area (SLA) level or, where this wasn't available, the child's postcode. One estimate is provided for each time point representing a linear interpolation of the data at the censuses either side of the time period. For example, if a SLA had 4.2% of people with ATSI origins in 2001 and 6.5% with ATSI origins in 2006 then the estimate for the proportion in 2004 would be:

$$estimate = 2001Data + (2006Data - 2001Data) \times \frac{time\_since\_census}{time\ between\ censuses},$$

$$estimate = 4.2\% + (6.5\% - 4.2\%) \times \frac{(2004 - 2001)}{(2006 - 2001)}$$

$$estimate = 4.2\% + 2.3\% \times 0.6$$

$$estimate = 5.6\%$$

If data is only available for one of the censuses then no interpolation is performed. A "'link type" variable is included to tell data users whether the linkage was performed using SLA or postcode and whether the 2001 census, 2006 census, 2011 census or all were used.

5.1.2 National Childcare Accreditation Council data

A key research question in LSAC relates to the impact of child care on children's developmental outcomes over time. While LSAC collected parent-report information on children's child care histories and carer reports on the child care environment, relatively little systematic information was collected on the quality of child care.

The National Childcare Accreditation Council Inc. (NCAC) has quality assurance data on every long day care (LDC) centre, some family day care (FDC) schemes and some before and after school care providers. The LSAC dataset includes linked NCAC data for most children using LDC or FDC at Wave 1, where contact details of this care were obtained and matched with NCAC data. The match rate obtained during the linkage process was 78% for Wave 1, 82% for Wave 2, 84% for Wave 3 and 92% for Wave 4.

One complication in using the NCAC data is due to the change of accreditation systems for both FDC and LDC. In Wave 1, all cases had FDC assessed under the guidelines laid out in 2nd edition of the*FDCQA Quality Practices Guide* (NCAC, 2004), while from Wave 2 and onwards all cases have been assessed under the 3rd edition of this guide, which was introduced in July 2005. The revised guidelines contain the same quality areas (though some have been combined) but have had the number of principles used to assess these areas reduced from 35 to 30. The old scheme had 10 quality areas assessed by 35 principles, while the new scheme has seven quality areas assessed by 30 principles.

For LDC, all Wave 1 centres were assessed under the*QIAS Validation Report*, 2nd Edition (NCAC, 2003). From July 2006, accreditation decisions were made under the*QIAS Quality Practices Guide*, 1st Edition. As a consequence, some of the Wave 2 and 3 accreditations were made under the new scheme, while some were made under the old scheme.

Before- and after-school care arrangements are assessed in the guidelines laid out in the*OSHCQA Quality Practices Guide*, 1st Edition (NCAC, 2003). In Waves 2 and 3, some accreditations were made under the new scheme, while some were made under the old scheme.

The variables included are:

- date of accreditation
- date of validation
- accreditation status
- LDC v1 Quality area 1: Relationships with children
- LDC v1 Quality area 2: Respect for children
- LDC v1 Quality area 3: Partnerships with families
- LDC v1 Quality area 4: Staff interactions
- LDC v1 Quality area 5: Planning and evaluation
- LDC v1 Quality area 6: Learning and development
- LDC v1 Quality area 7: Protective care
- LDC v1 Quality area 8: Health
- LDC v1 Quality area 9: Safety
- LDC v1 Quality area 10: Managing to support quality
- LDC v2 Quality area 1: Staff relationships with children and peers
- LDC v2 Quality area 2: Partnerships with families
- LDC v2 Quality area 3: Programming and evaluation
- LDC v2 Quality area 4: Children’s experiences and learning
- LDC v2 Quality area 5: Protective care and safety
- LDC v2 Quality area 6: Health, nutrition and wellbeing
- LDC v2 Quality area 7: Managing to support quality
- FDC Quality area 1: Interactions
- FDC Quality area 2: Physical environment
- FDC Quality area 3: Children's experiences, learning and development
- FDC Quality area 4: Health, hygiene, nutrition, safety and wellbeing
- FDC Quality area 5: Carers and coordination unit staff
- FDC Quality area 6: Management and administration
- OHS Quality area 1: Respect for children
- OHS Quality area 2: Staff interactions and relationships with children
- OHS Quality area 3: Partnerships with families and community links
- OHS Quality area 4: Programming and evaluation

- OHS Quality area 5: Play and development
- OHS Quality area 6: Health, nutrition and wellbeing
- OHS Quality area 7:Protective care and safety
- OHS Quality area 8: Managing to support quality
- Demographic data

The data used to develop the quality areas was collected from six sources:

- a self-study report prepared by centre management
- a validation survey completed by the director
- a validation survey completed by staff
- a validation survey completed by families
- a validation report completed by an independent peer
- a set of moderation ratings completed by independent moderators.

Data on 35 principles were collected. Each principle was related to one of the 10 quality areas. Response categories for each principle were: “unsatisfactory”, “satisfactory”, “good quality” and “high quality”. Proportionally weighted factor-score regression coefficients for principle ratings were calculated to determine the extent to which each principle contributed to a quality area. For further information, see Rowe (2006).

As no data about the child was obtained, no consent was required from parents to collect this data (although parents did need to give details about their carers to assist in the linking).

5.2 Supplementary files

5.2.1 Time use diary data

In Waves 1 to 3, responding families were given two time use diaries (TUDs) to complete at each wave. Each record in the TUD data relates to a single diary; that is, each child can have up to two records (one for each TUD).

The key component of the TUD data is to gather information on children’s activities and context for the 96 15-minute periods of each 24-hour block. In addition to these variables, the TUD data includes the child’s unique identification number in order to allow linkage with the main dataset. It also includes the following general descriptors:

- date diary should be completed
- day of week diary should be completed
- diet of the study child on the day in question (Waves 2 and 3)
- relationship of the diary writer to the child
- over what duration the diary was completed
- actual day and date of completion
- hours of work done by respondent on day of completion (Waves 2 and 3)
- the kind of day described in the diary.

Due to scanning problems in Wave 1, and other data quality issues that are likely to apply equally across waves, a number of imputations and corrections have been applied to the TUD data (see *Data Issues Waves 1 to 6* for details, available at <www.growingupinaustralia.gov.au/pubs/issues/index.html>). So researchers can determine the effect of these imputations/corrections to the data on any analysis. An uncorrected version of the TUD data is also provided, as well as files containing imputations/corrected versions of cases that were considered unsuitable for data analysis even after correction.

LSAC Technical Papers 4 and 13 include detailed discussions of issues that should be considered when using the time use data. The technical papers are available from <www.aifs.gov.au/growingup/pubs/technical/index.html>.

In Wave 4 a new methodological approach was undertaken. The study shifted from the parent being the informant to the study child being the informant. In Waves 4 to 6 only the K cohort completed the TUD, which was substantially different from the TUDs that the parents completed in earlier waves. The TUD in Waves 4 to 6 had the form of an “ABS Activity Episode” diary. This data is stored as a long file, as opposed to the wide files the previous diaries were stored as.

Example analysis

SAS

The following code gives the proportion of children eating or drinking while watching a TV, video, DVD or movie at any time of day for the B cohort at Wave 1. Statements 1 and 2 tell SAS to create a new dataset beginning with the data in the mtud.diary2 file (you will need to use your own library name). The third statement tells SAS to treat the time use data as a multidimensional array (x) containing 96 rows of 40 columns each. The next statement tells SAS to set up a new array of 96 variables (TVeat) into which the data for eating in front of the TV will be derived.

Statements 5 to 8 contain a do loop, which runs across all 96 time periods. Statement 5 tells SAS to create a variable “i” to keep track of which time period is being worked on, and to give it the values 1 to 96 in turn. Statement 6 tells SAS to allocate the value 100 at the position in the “TVeat” array for the current time period if the child was eating or drinking (column 4 in the array “x”) and was watching a TV, etc. (column 12 in “x”). Statement 7 says the value of 0 will be assigned if the child either wasn’t eating or drinking or wasn’t watching TV, etc. and the diarist wasn’t unsure of the child’s activities for the time period. This means that cases where the diarist wasn’t sure, or didn’t fill any information in for activities in this time period, will have missing data. Statement 8 finishes to do loop, and statement 9 finishes the data step so SAS runs the above statements.

Statements 10-13 produce the means of the variables in the “TVeat” array (which SAS gives the names TVeat1 to TVeat96 by default). The mean here will be the percentage of children from whom an activity was known that ate or drank in front of the TV, etc. at each time period. Line 12 uses the day weight variable “bweightd” to ensure the proportion is representative of the population and represents each day of the week equally.

```
data diary2;
        set mtud.diary2;
        array x [96,40] b2da0101--b2de0196;
        array Tveat [96];
        do i=1 to 96;
                if x[i,4]=1 and x[i,12]=1 then Tveat[i]=100;
                else if (x[i,4]=0 or x[i,12]=0) and x[i,1]^=1 then Tveat[i]=0;
        end;
run;
proc means data=diary2;
        var Tveat1-Tveat96;
        weight bweightd;
run;
```

This data can be used to produce a graph known as a tempogram.

Figure 2 shows the data produced by the example program along with the equivalent data for the K cohort at Waves 1 and 2. It shows that children did more of this as they got older, and that this activity was most common in the early mornings.

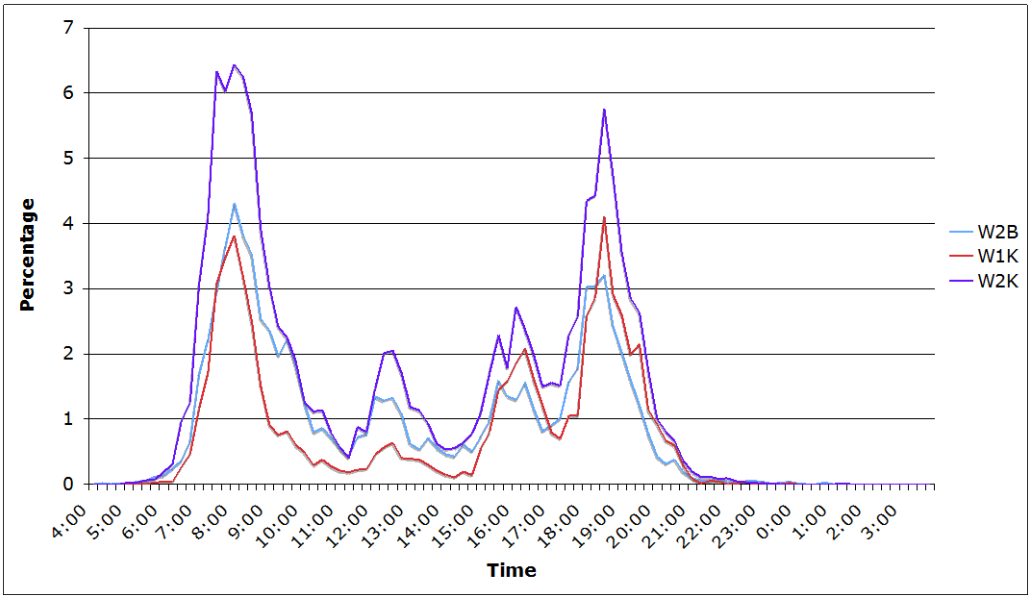


Figure 2: Tempogram of children watching TV, video, DVD or movie while eating or drinking by wave and cohort.

SPSS

The equivalent code to derive the TVeat variable in SPSS would appear as:

```
do repeat
eat b2da0401 b2da0402 … b2da0496/
tv b2da1201 b2da1201 … b2da1296/
dk b2da0101 b2da0101 … b2da0196/
tve tveat1 to tveat96.
if (eat=1 or tv=1) tve=1.
if ((eat=0 or tv=0) and dk=0) tve=0.
end repeat.
```

STATA

The equivalent code to derive the TVeat variable in STATA would look like

```
foreach n of numlist 1/9 {
gen tveat`n'=1 if (b2da040`n'==1 & b2da120`n'==1)
replace tveat`n'=0 if ((b2da040`n'==0 | b2da120`n'==0) & b2da010`n'==0)
}
foreach n of numlist 10/96 {
gen tveat`n'=1 if (b2da04`n'==1 & b2da12`n'==1)
replace tveat`n'=0 if ((b2da04`n'==0 | b2da12`n'==0) & b2da01`n'==0)
}
```

5.2.2 Medicare Australia data

In Wave 1, 97% of parents of study children gave consent for their children's data to be linked with Medicare Australia data for the duration of the study. This includes data from the Medicare Benefit Scheme (MBS), the Pharmaceutical Benefit Scheme (PBS) and the Australian Childhood Immunisation Records (ACIR). Data from

these sources provide an indication of usage history of MBS, PBS and ACIR services. Linkage was successful for 93% of children (incomplete consent forms resulted in data not being released for about 400 children).

Since the child's use of medical services is ongoing, the Medicare Australia data are not broken into waves, but are provided as three separate files:

ACIR: Each record in the file represents an immunisation that the child has had.

MBS: Each record on this file represents a benefit claim.

PBS: Each record represents a benefit claim.

ACIR file

Records are currently available for payments received from birth to early 2013. The following variables are included on the file:

- child identification number
- vaccination code
- vaccination name
- scrambled provider ID
- date of receipt of payment
- date of immunisation.

Some of the vaccination codes contain dose numbers, indicating a vaccine that has been received in a series of doses. The sequence of doses for these has been included in the dataset (i.e. 1st, 2nd, etc.). If a dose is missing, it means that it was either not reported to ACIR or it was missed.

MBS file

Records are currently available for services between January 2002 (or birth for the B cohort) and early 2015. The following variables are included on this file:

- child identification number
- item number
- item name
- amount of benefit paid
- hospital indicator
- scrambled provider ID
- date of payment
- date of service.

Some cases have very small or negative benefit amounts. In relation to negative benefits, this indicates that an adjustment has been made to the Medicare benefit records. There are several reasons why this may happen:

- It is a correction of a data entry made against the wrong individual reference number on a Medicare card (i.e., service is initially incorrectly recorded against someone else on the same card).
- The provider has issued an amended account.
- A new cheque has been issued to replace lost/stolen/unpresented cheques.

In relation to small benefits:

- There are a number of item numbers that have small benefits; for example, many pathology-related claims.
- There are also small amounts for things such as bulk bill incentives (generally around $5-6).
- The claimant had reached the Medicare Safety Net (MSN) threshold. Once the threshold has been reached, the family's out-of-pocket expenses are tallied and a payment is calculated for a percentage of the substantiated

amounts. In effect, there can be two payments made for the same doctor's visit - one to the doctor for the service and one to the claimant for MSN purposes.

PBS file

The final of these datasets contains the PBS data. Again, each record represents a benefit claim. Records are available for medications supplied between May 2002 (or birth for the B cohort) and early 2015. The following information is included for each record:

- child identification number
- item code
- item name
- quantity
- benefit paid
- prescription type (original, repeat or unknown)
- payment category
- payment status
- date of payment
- date of supply.

Example derivations

There are simple techniques in SAS, SPSS and STATA to summarise across multiple records to create derived items from the Medicare datasets. The following code samples create a variable (ben07) for the amount of PBS benefits paid for a child in 2007. Note that this variable will initially be missing for cases that had no PBS claims in 2007 as well as those for which data linkage was unsuccessful. The ”match” file can be used to distinguish between these cases and set ben07 to 0 for those with no claims. This file contains a variable called ”medicare”, which is 1 if linkage is successful for a case and 0 otherwise.

SAS

```
proc means data=m.pbs nway sum;
        class hicid;
        var benefit;
        where datesupp>=mdy(1,1,2007) and datesupp<=mdy(1,1,2008);
        output out=temp sum=ben07;
run;
data temp;
        merge temp m3.match;
        by hicid;

        if medicare=1 and ben07=. then ben07=0;
run;
```

SPSS

```
temp.
select if (datesupp >= date.dmy(1,1,2007) & datesupp <= date.dmy(31,12,2007)).
aggregate
```

```
  /outfile='/temp.sav'
  /break=hicid
  /ben07=sum(benefit).
get
  file='/temp.sav'.
match files /file=*
  /file='/match.sav'
  /by hicid.
if  (medicare=1 & missing(ben07)) ben07=0.
execute.
```

STATA

Note that the collapse command will delete all other data than hicid and ben07. Ensure it is saved to a new file.

```
collapse (sum) ben07=benefit if (datesupp>=mdy(1,1,2007) & datesupp<=mdy(1,1,2008)), by(hicid)
merge hicid using match
replace ben07=0 if (medicare==1 & ben07==.)
keep if ben07!=.
sort hicid
save temp, replace
```

5.2.3 Household composition data

At each wave of data collection, responding families are asked to give the details of the people currently residing in their household, as well as people who have come and gone between waves but lived with the study child for at least three months.

This dataset contains one record for each study child, detailing the composition of their household from their recruitment to the study to the most recent data collection.

Details collected about the study child, Parent 1 and Parent 2 are included in each main dataset, along with a number of derived variables on household composition.

5.2.4 LSAC NAPLAN data

In Wave 3, 81% of parents of K cohort children gave consent for their child's data to be linked with NAPLAN data for the duration of the study. Linkage was successful for 96% of children. For 4% of children, the NAPLAN data were not found, either because these children had not sat NAPLAN tests yet or they sat the NAPLAN tests in 2008 or 2009 but a match was not found. Families who did not give consent or who did not participate at Wave 3 were asked again at Wave 4. Out of 964 families who were followed up in Wave 4, 847 gave consent to link NAPLAN results.

The Wave 6 LSAC NAPLAN release includes B cohort & K cohort NAPLAN results for 2008 to 2014.

LSAC Technical Paper 8 includes a detailed discussion of data compendium and data issues that should be considered when using the LSAC NAPLAN data. The report is available from www.aifs.gov.au/growingup/pubs/technical/index.html.

5.2.5 Wave 2.5 data

The data from the Wave 2.5 mail-out is included in two separate datasets. Unlike Wave 1.5 in relation to Wave 1, families that responded to Wave 2.5 did not necessarily respond to Wave 2. Merging these with the Wave 2 datasets would have resulted in a number of largely blank cases on the data file.

The data in the Wave 2.5 file consists of questionnaire items, a small number of derived items and linked census data based on the postcodes of responding families at the time of Wave 2.5. Unfortunately, formatting of the questionnaires resulted in some respondents skipping items they should have answered. Imputation has been performed on some items where it was possible to infer the data for these questions based on responses to other questions. See *Data Issues Waves 1 to 6* for further information (available from www.growingupinaustralia.gov.au/pubs/issues/index.html).

5.2.6 Wave 3.5 data

The data from the Wave 3.5 mail-out is included in a separate dataset, in the same way that data from Wave 2.5 was included.

The data in the Wave 3.5 file consists of questionnaire items, a small number of derived items and linked census data based on the postcodes of responding families at the time of Wave 3.5. Imputation has been performed on some items where it was possible to infer the data for these questions based on responses to other questions. . See *Data Issues Waves 1 to 6* for further information (available from www.growingupinaustralia.gov.au/pubs/issues/index.html).

5.2.7 ACARA MySchool data

Data has been obtained from ACARA. ACARA is responsible for collating NAPLAN data received from Australian schools, collecting school characteristics and managing the MySchool website. Some of the data ACARA collects and collates on Australian schools is publicly available on the MySchool website. School data about the schools LSAC participants attend has been linked onto the LSAC survey datasets and is available to data users.

5.2.8 AEDC data

Every 3 years since 2009, the Australian Government has undertaken a census of all children in their first year of full-time schooling. This information is used to identify communities where families and children may require extra support and to help shape the future and wellbeing of Australian children. (www.aedc.gov.au/parents)

Data has been obtained from the Department of Education. The Department of Education is responsible for the Australian Early Development Census. The Social Research Centre manages the data. The data contains no variable labels or value labels, but these can be found in the Data Dictionary provided and on the AEDC web site. See www.aedc.gov.au and www.aedc.gov.au/researchers/resources-for-researchers/data-dictionary

6 Variable naming conventions

The variable naming convention was developed so that variables have predictable names across waves and informants, and so that thematically linked variables have similar names wherever possible. A two-page “help sheet” is included on the LSAC Data CD to help users learn these conventions.

6.1 Questionnaire variables

Variable names follow the standard format in most cases. Exceptions to this naming convention (derived items and household composition variables) are explained in the sections that follow.

Standard format: A tt xxxxx

Where:

A = child age indicator

tt = topic indicator

xxxxx = specific question identifier.

6.1.1 Child age indicator (alpha)

The child age indicator is an alpha symbol that indicates the child’s age, allowing for comparisons between the cohorts where data has been collected for both cohorts at that age. For instance:

a indicates the child is aged 0-1 years (which is the B cohort in Wave 1)

b indicates the child is aged 2-3 years (which is the B cohort in Wave 2)

c indicates the child is aged 4-5 years (which is the B cohort in Wave 3, and the K cohort in Wave 1)

d indicates the child is aged 6-7 years (which is the B cohort in Wave 4, and the K cohort in Wave 2)

e indicates the child is aged 8-9 years (which is the B cohort in Wave 5, and the K cohort in Wave 3)

f indicates the child is aged 10-11 years (which is the B cohort in Wave 6, and the K cohort in Wave 4)

g indicates the child is aged 12-13 years (which is the K cohort in Wave 5)

h indicates the child is aged 14-15 years (which is the K cohort in Wave 6).

This is an example of how the child age indicator is used for the item 'Parent 1 rating of parenting self-efficacy':

Wave 1 B cohort: apa01a

Wave 2 B cohort: bpa01a

Wave 3 B cohort: cpa01a

Wave 4 B cohort: dpa01a

Wave 5 B cohort: epa01a

Wave 6 B cohort: fpa01a

Wave 1 K cohort: cpa01a

Wave 2 K cohort: dpa01a

Wave 3 K cohort: epa01a

Wave 4 K cohort: fpa01a

Wave 5 K cohort: gpa01a

Wave 6 K cohort: hpa01a

Those items of information that do not change (e.g., details of birth, age child began or stopped something, etc.) are given the age indicator *z* so that they have a consistent variable name across cohorts regardless of the age of the child when the information was obtained. For example, zhs03a indicates “birth weight of the study child”

regardless of whether the information was collected when the child was aged 0-1 years as for the B cohort, or aged 4-5 years as for the K cohort.

6.1.2 Topic indicator (alpha)

The topic indicator is taken from the topic field of the data dictionary. Efforts were made to make the abbreviations used meaningful (e.g., family demographics is fd).

A list of topics and their abbreviations is provided in Table 4.

Table 4: Topics used in LSAC datasets

Abbreviation	Topic	Scope
fd	Family demographics	Demographic information relating to the family such as education, ethnicity and religion
fn	Finances	Financial information such as income and use of government benefits
gd	General development	Scales that contain items from multiple domains of child development
hb	Health behaviour and risk factors	Behaviours and other risk factors that potentially impinge upon the health of the study child or his/her family. These include behaviours such as parental smoking and drinking as well as risk factors such as a parent experiencing diabetes during pregnancy.
he	Home education environment	Information on factors likely to impinge on the child's learning while at home such as parental support for education, number of books in the home and TV use. Also contains information on parent interaction with teachers such as parent teacher interviews including from the teacher's perspective.
ho	Housing	Information on housing such as number of bedrooms, tenure type and payments
hs	Health status	Information about the physical and mental health status of the study child or his/her family such as body mass index, diagnosis of conditions and number of hospital stays
id	Identifiers	Questionnaire process variables such as sequence guides, consents and details of proxy respondents
lc	Learning and cognition outcomes	Information on the child's development in the areas of learning and cognition including language, literacy and numeracy
pa	Parenting	Information on parenting styles and other information affecting parenting such as self-efficacy
pc	Program characteristics	Characteristics of the educational or childcare program such as type of program, number of days or hours the child attends and staff satisfaction
pe	Parent living elsewhere	Details of the child's PLE such as the relationship to study child, interactions with resident parents and child support
pl	Parental leave in Australia	Data from the" "Parental Leave in Australia" Survey—a nested study
pw	Paid work	Information on work status such as employment, occupation and work/family interactions
re	Relationships	Information on the quality of relationships primarily focused on the relationship between Parent 1 and Parent 2, but also on broader family harmony
sc	Social capital	Information on social capital such as attitudes to neighbours and the neighbourhood and use of services
se	Social and emotional outcomes	Information relevant to the social and emotional development of the child such as temperament, behaviour and emotional states

tp	Teaching practices	Practices employed by teachers and childcare workers in their work such as time use, use of resources and general philosophies

For example:

apa01a (Parent 1 rating of self-efficacy) has “pa” as the second and third letters as its topic is “Parenting”; and

zhs03a (Birth weight of study child) has “hs” as the second and third letter as its topic is “Health status”.

6.1.3 Specific question identifier (alphanumeric)

The last five digits of a variable name make up the specific question identifier (if required). These digits contain whatever information is necessary to uniquely identify each item. Each has an arbitrary two-digit question number, not related to the questionnaire positioning. Items of related content are grouped together as much as possible.

For example:

bhs12a is whether Parent 1 is concerned about the child’s weight.

bhs12b is whether Parent 1 considers the child to be “underweight”, “normal weight”, “somewhat overweight” or “very overweight”.

The sixth digit of the variable name can also be an informant or subject indicator where a question is asked of or about more than one person. The indicators used are:

a Parent 1

b Parent 2

c Study child

m Mother

f Father (or family home for census data)

t Teacher/Carer

i In-between waves respondent

For example:

bhs13a is Parent 1’s rating of their own overall health status.

bhs13b is Parent 2’s rating of their own overall health status.

bhs13c is Parent 1’s rating of the study child’s overall health status.

bhs13p is the PLE’s rating of their own overall health status.

bhs13m is the mother’s rating of their own overall health status.

bhs13f is the father’s rating of their own overall health status.

An exception to the above rule is in the area of childcare and education (variables with topic indicators pc and tp). Here the prefixes a, b, c, d and e are used to mean different things at each wave depending on the options available to the child at that age (see Table 5).

Table 5: Subject indicators for education and childcare variables				
Indicator	Age 0-1	Age 2-3	Age 4-5	Age 6-7
a	1st child care	1st child care	Main educational program	Main educational program
b	2nd child care	2nd child care	1st child care	Before-school care
c	3rd child care	3rd child care	2nd child care	After-school care
d		Other child care	3rd child care	
e				Program child would attend if attending school
o		Any extra care	Any extra care	Any extra care
Indicator	Age 8-9	Age 10-11	Age 12-13	Age 14-15
a	Main educational program	Main educational program	Main educational program	Main educational program
b	Before-school care	Before-school care	Before-school care	
c	After-school care	After-school care	After-school care	
d	Child care at other times		Other child care	Other child care
e	Program child would attend if attending school	Program child would attend if attending school		
o		Any extra care		

All items that form a scale have a single question number. Where applicable, the name of the item also indicates the relevant subscale or sub-subscale (please note that this is done only where it is possible to do so, due to the eight character limit for the name of an item).

An example of how this is applied is shown with the Conduct Problems and Peer Problems subscales of the Strengths and Difficulties Questionnaire (see Table 6). These are subscales that both Parent 1 and the teacher filled out in Waves 1 and 2 for the K cohort.

As shown:

- The 6th character in the variable name in this case represents an informant indicator: “a” is for Parent 1, “t” is for teacher.
- The 7th character indicates the subscale: 4 for Conduct, 5 for Peer. (Note: the subscales 1 for Prosocial, 2 for Hyperactivity and 3 for Emotional are also available as part of the SDQ.)
- The final character uniquely identifies each item. (Note: different items were used for the Conduct subscale in Waves 1 and 2 due to the change in the child’s age).

Table 6: Variable names of SDQ[a] conduct and peer problems subscales				
	Wave 1 Parent 1 K cohort name	Wave 1 Teacher K cohort name	Wave 2 Parent 1 K cohort name	Wave 2 Teacher K cohort name
Conduct problems				
Often loses temper	*cse03a4a*	*cse03t4a*	*dse03a4a*	*dse03t4a*
Generally well behaved, usually does what adults request	*cse03a4b*	*cse03t4b*	*dse03a4b*	*dse03t4b*
Often fights with other children or bullies them	*cse03a4c*	*cse03t4c*	*dse03a4c*	*dse03t4c*
Often argumentative with adults	*cse03a4d*	*cse03t4d*	*N/A*	*N/A*

Can be spiteful to others	*cse03a4e*	*cse03t4e*	*N/A*	*N/A*
Often lies or cheats	*N/A*	*N/A*	*dse03a4f*	*dse03t4f*
Steals from home, school or elsewhere	*N/A*	*N/A*	*dse03a4g*	*dse03t4g*
Peer problems				
Rather solitary, tends to play alone	*cse03a5a*	*cse03t5a*	*dse03a5a*	*dse03t5a*
Has at least one good friend	*cse03a5b*	*cse03t5b*	*dse03a5b*	*dse03t5b*
Generally liked by other children	*cse03a5c*	*cse03t5c*	*dse03a5c*	*dse03t5c*
Picked on or bullied by other children	*cse03a5d*	*cse03t5d*	*dse03a5d*	*dse03t5d*
Gets on better with adults than with other children	*cse03a5e*	*cse03t5e*	*dse03a5e*	*dse03t5e*

Note: [a]The SDQ is copyrighted by Robert Goodman, UK, 1999.

6.2 Derived variables

The derived items start with an age indicator, as outlined in section 6.1.1, followed by an informant or subject indicator and then a mnemonic that relates to the subject matter of the derived item. For example, the Peer subscale of the SDQ for the K cohort teacher in Wave 2 is dtpeer, where d = child aged 6-7 years, t = teacher and peer = Peer subscale of SDQ.

6.3 Study child household composition variables

In order to keep the variable names under eight characters, it was necessary to have a slightly different convention in the Wave 2 data release. Household composition variables have the following structure:

A f ##xmmm

Where:

A = Child age indicator

f = f (for “family”)

= Question number (numeric)

x = Sub-question indicator (optional)

mmm = person identifier

Note:

The age indicator above is as described in section 6.1.1.

“f” is a constant to indicate that it is the household composition that is being described.

The question number and sub-question indicator indicate the question being responded to.

The person identifier indicates the member number, or other identification information. For every household, the study child is Member 1, the Wave 1 Parent 1 is Member 2, and the Wave 1 Parent 2 is Member 3 (or will be missing if there is no Parent 2 at Wave 1). Any additional people in the household at the time of Wave 1 are given member numbers 4 through to whatever is required. Each household member retains the same member number throughout the study, even if they leave and re-enter the study child’s home.

Due to the requirements of the CAI instrument, some families have “gaps” in member numbering; for example, where someone is Member 5 but Member 4 has never been assigned.

Member 1 is denoted by “m1” in the above convention, Member 2 as “m2” and so on as required.

As families change from Wave 2 on, the new Parent 1, Parent 2, Mother or Father could have any member number apart from 1. For this reason an extra set of variables has been derived to give the details for the Parent 1, Parent 2, Mother and Father at any age. This subscript is an age indicator and then either “p1”, “p2”, “m”, or “f”.

A set of indicator variables tracks the household member number of Parent 1, Parent 2, Mother and Father at each wave. For example, bp2mn tells you the household member number of Parent 2 when the child is aged 2-3, while cmmn gives the member number of the mother when the child is aged 4-5.

Some examples:

> zf02m1 is the gender of the study child (z = unchanging characteristic, f = “Family”, 02 = gender, m1 = study child)
>
> bf01m2 is whether the Wave 1 Parent 1 is present in the household when the child is aged 2-3 (b = child aged 2-3, f = “Family”, 01 = present for wave, m2 = Wave 1 Parent 1)
>
> cf01m3 is whether the Wave 1 Parent 2 is present when the child was aged 4-5 (or whether there was a Parent 2 at all in Wave 1 for the K cohort) (c = child aged 4-5, f = “family”, 01 = present for wave, m3 = Wave 1 Parent 2)
>
> af08am is the relationship of the Mother to the study child when the child was aged 0-1 (a = ages0-1, f = “family”, 08 = relationship to study child, am = mother of child at age 0-1)
>
> df01cp1 is whether the Parent 1 of the child when aged 4-5 is present in the household when the child is aged 6-7. (d = child aged 6-7, f = “family”, 01 = present for wave, cp1 = child’s Parent 1 when child is aged 4-5)
>
> cf13dp2 is whether the Parent 2 of the child when aged 6-7 had a medical condition or disability at the time the child was 4-5 (c = child aged 4-5, f = “family”, 13 = whether person has a disability, dp2 = Parent 2 when child is aged 6-7).

Table 7 shows the information that is available for each person.

Table 7: Question numbers used in variable names for household member characteristics

##x	Question
01	Present for wave
02	Gender
03	Age
04	Date of birth
05	Temporarily away from home (as per Wave 1 question)
06	Relationship to Parent 1
07	Relationship to Parent 2
08	Relationship to study child
09	Country of birth
10	Year of first arrival in Australia
11	Language other than English spoken at home
12	ATSI status
13	Has a condition or disability for 6 months or more (as per Wave 1 question)
13a	1st specific condition
13b	2nd specific condition
14	Date stopped living with study child
15	Reason stopped living with study child
16	Temporarily away from home (as per Wave 2 question)
16o	Temporarily away from home (other) (as per Wave 2 question)
17	Has a condition or disability for 6 months or more (as per Wave 2 question)
17a	Has sight problems (as per Wave 2 question)
17b	Has hearing problems (as per Wave 2 question)
17c	Has speech problems (as per Wave 2 question)
17d	Has blackouts, etc. (as per Wave 2 question)
17e	Has difficulty learning (as per Wave 2 question)
17f	Limited use of arms or fingers (as per Wave 2 question)
17g	Difficulty gripping (as per Wave 2 question)
17h	Limited use of legs and feet (as per Wave 2 question)
17i	Other physical condition (as per Wave 2 question)
17j	Other disfigurement (as per Wave 2 question)
17k	None of the above conditions (as per Wave 2 question)
18	Restricted in everyday activities
18a	Has difficulty breathing (as per Wave 2 question)
18b	Has chronic pain (as per Wave 2 question)
18c	Has nervous condition requiring treatment (as per Wave 2 question)
18d	Has mental illness requiring supervision (as per Wave 2 question)
18e	Has head injury (as per Wave 2 question)
18f	Has other long-term condition (as per Wave 2 question)
18g	Has other condition requiring treatment (as per Wave 2 question)
18h	None of the above restrictions (as per Wave 2 question)
19	Date began living with the study child
20	Household member was in the household for at least 3 months, but moved in and left between current and previous waves
21	Person type

6.4 PLE household composition variables

From Wave 4, the household information for the child's parent living elsewhere (PLE) has been collected. PLE household composition variables have a similar structure to that of the study child household composition variables:

A f ##xple#

Where:

A = Child age indicator

f = f (for “family”)

= Question number (numeric)

x = Sub-question indicator (optional)

ple# = person identifier within PLE household with ple (for Parent Living Elsewhere) and # member number

Note:

The age indicator is as described in section 6.1.1.

“f” is a constant to indicate that it is the household composition that is being described.

The question number and sub-question indicator indicate the question being responded to.

The person identifier comprises the constant “ple” to indicate that it is the PLE household and the member number. For every PLE household, **the study child is Member 1 (ple1)** and **PLE is Member 2 (ple2)**. For example, variable f02ple2 refers to a PLE gender when a study child is 10-11 years old. Any additional member in the household is assigned a PLE member number that remains the same throughout the study, even if they leave and re-enter the PLE’s home.

Table 8 shows the information that is available for each PLE.

Table 8: Question numbers used in variable names for PLE household member characteristics

##x	Question
01	Present for wave
02	Gender
03	Age
04	Date of birth
05	Temporarily away from home (as per Wave 1 question)
06a	Relationship to PLE
08	Relationship to study child
09	Country of birth
10	Year of first arrival in Australia
11	Main language spoken at home
12	ATSI status

A PLE household file also includes the following variables (the asterisk refers to the child age indicator):

*datplec—date of PLE CATI interview

*plepar—whether PLE has a partner

*pleparmn—PLE partner member number in PLE household

*dfd02p3—date of recent PLE marriage

*dfd02p4—date of PLE cohabitation.

6.5 Age invariant indicator variables

There are five variables at the start of each of the main data files that contain no age indicator. These are:

hicid—unique identifier assigned when child was selected by Medicare Australia

cohort

wave

stratum—stratum at the time of selection

pcodes—postcode at the time of selection

Users wishing to create long datasets should note the presence of these variables when removing age indicators.

6.5.1 Study child unique identifier

Each study child has a single, unique identification variable to ensure matching and merging across instruments, files and waves. This number was allocated at the time of selection by Medicare Australia.

The first digit indicates which cohort the child is in (1-4 = Infant; 5-8 = Child) and what fieldwork phase (see "Methodology" section for more detail) the child was selected to be part of in Wave 1 (Phase 1 = 1 and 5, Phase 2 = 2 and 6, etc.).

The second is the state the child was selected from (1 = NSW, 2 = Vic., etc.).

The third indicates the part of state the child was selected from (1-2 = capital city; 3-4 = rest of state).

The remaining five digits are a random number allocated by Medicare Australia.

Note that the stratum for selection may differ from the location of the child at interview and that the fieldwork phase may change from wave to wave.

6.6 Indicator variables

There are indicator variables in the main data files that indicate which parts of an interview were incomplete. These variables were created to flag to data users (through yes/no values) that no data, or only partial data, exists for an instrument (for example the CASI) or an informant (e.g., parent 1). The data may be incomplete due a number of different reasons. There may be no data if a self-complete form was not returned; parent/child did not provide consent to obtain/provide the data; one of the informants refused to participate; or when the interview was only partially completed.

For example, on the day of the interview the parent may consent to the child participating but refuse to participate themselves. In this example, there would be data for the sections where the study child is the informant; however, there would be no data for the sections where Parent 1 is the informant. To identify these cases a data user can use the following indicator variable *__nopar (*__ refers to the age indicator). Another example is a teacher's responses. To identify cases where a teacher form was not returned, a data user can examine the variable ***tcd**. A data user can also examine the following indicator variables: ***partresp** to identify cases that were incomplete due to an interview stopping half way as opposed to just certain sections being refused or ***hhresp** to identify cases where the household interview was completed.

There are a large number of indicator variables and data users are encouraged to investigate the reasons for data being incomplete through these variables. Note that the indicator variables do not follow the general variable naming conventions described above. Some indicator variables are listed in Table 1. Indicator variables can be found in the data dictionary under the topic "Identifiers", along with other variables that fall under that topic. For more information refer to the data dictionary.

6.7 Variable labelling convention

The labels used for the variable dataset take the following general form:

(Age) - (Informant/subject) - (Questionnaire position) - (Construct label)

Age is a label for the age indicator from the variable name, so:

a = 0/1

b = 2/3

c = 4/5

d = 6/7

e = 8/9

f = 10/11

g = 12/13

h = 14/15

If no age indicator is present in the variable name, or the age indicator is z, then this part of the variable label will not be included.

For example:

label zf04m1 = "SC - DOB", here no age is associated with the variable because it doesn't change with time, hence no age indicator is included.

label df03m1 = "6/7 - SC - Age", this variable is a variable that changes over time so the age indicator is required in order to establish when the question was answered.

Informant/subject gives the informant or subject of the question as contained in the variable name. For household composition variables involving Parent 1, Parent 2, Mother or Father, the age of the study child at which the person's status as parent is determined will also be indicated (e.g., M@0/1 is the Mother when the child is aged 0-1 years old). If the information only exists for one subject or informant in the study this part of the variable label will not be included.

Questionnaire position indicates the location of the question the data was obtained from within the LSAC questionnaires (e.g., F2F H2 is question H2 of the face-to-face interview). This part of the variable label is left blank for derived items such as scales and other non-input items but included for Mother/Father variables where the location of both the P1 and the P2 variables are given.

Construct label provides a description of what information is actually contained in the variable (e.g., "Sex", "Birthweight", etc.). This part of the variable name will be consistent for each variable representing the same construct for a different subject/informant or wave.

For example:

The Parent 1's rating of their own health quality at Wave 1 for the B cohort (ahs13a) has the variable label "0/1 - P1 - P1L D1 - Global Health Measure". (0/1 is the age indicator, P1 is the informant/subject indicator, P1L D1 indicates the variable comes from the first question of section D of the Parent 1 leave-behind questionnaire, "Global Health Measures" is the construct label).

Total score for the Parent 1 parental warmth scale for the K cohort at Wave 2 (dbwarm) id "6/7 – P2 – Warm parenting" (6/7 is the age indicator, P2 is the informant indicator, there is no questionnaire position as the variable is calculated from multiple questions, "Warm parenting" is the construct label).

6.8 Missing value conventions

Missing data are coded as follows:

-1 Not applicable (when explicitly available as an option in the questionnaire)

-2 Don't know

-3 Refused or not answered

-4 Section refused

-9 Not asked due to one of the following reasons:

(a) A question was skipped due to the answer to a preceding question (e.g., if a child never repeated a grade, the following question regarding what grade the child repeated was not asked/skipped).

(b) A form was not returned or consent to participate was not given (e.g., if a

teacher form was not returned then the teacher's responses for this hicid are set to -9. To identify cases for which a form was not returned/or consent was not provided a data user can use an indicator variable (see Table 1 for details)).

(c) One of the informants refused to participate (e.g., if a parent refused to participate but not a child then the parent's responses are set to -9. To identify cases when the parent refused to participate, a data user can use the *nopar indicator variable).

(d) A form was partially completed (e.g., Parent 1 completed the interview over the phone (P1 CATI) but the face-to-face component did not occur. To identify these cases, a data user can use the *partresp indicator variable). (See section 6.6 for more detail.)

-99 Negative income (loss)

. Missing data—data not collected where it might be expected (e.g., the respondent skipped a question they should have answered in a self-complete form), or made missing due to an unreliable value (e.g., weight of Parent 1 recorded as 800 kg).

For further details about how missing LSAC income data is treated see Technical Paper No. 14 available at <www.growingupinaustralia.gov.au/pubs/technical/index.html>.

7 Documentation

A number of tools can be used to navigate the LSAC dataset:

- instruments
- frequencies
- online LSAC data dictionary
- Excel spreadsheets of the data dictionary (good for creating hard copies)

Users should also consider which documents they want to print out and which they want to look at electronically. We have found that the marked-up questionnaires and interview specifications are best printed and provide the easiest method of browsing to familiarise yourself with the data available. The data dictionary is best used for searching for specific items and mapping items from wave to wave.

These tools are described in more detail below.

7.1 Marked-up instruments

The associated variable name has been added beside each question in the questionnaires and interview specifications. Derived variables are also included. See Figure 3 for an example.

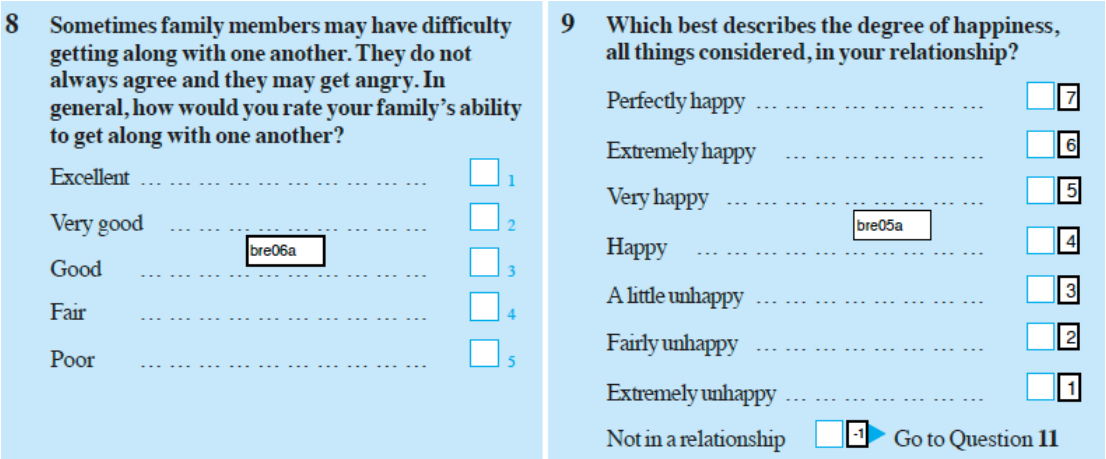


Figure 3: Example of the marked-up questionnaires.

A mock questionnaire (interview specifications) has also been generated for the CASI and CAI instruments used in Waves 2 to 6. Figure 4 is a sample of this.

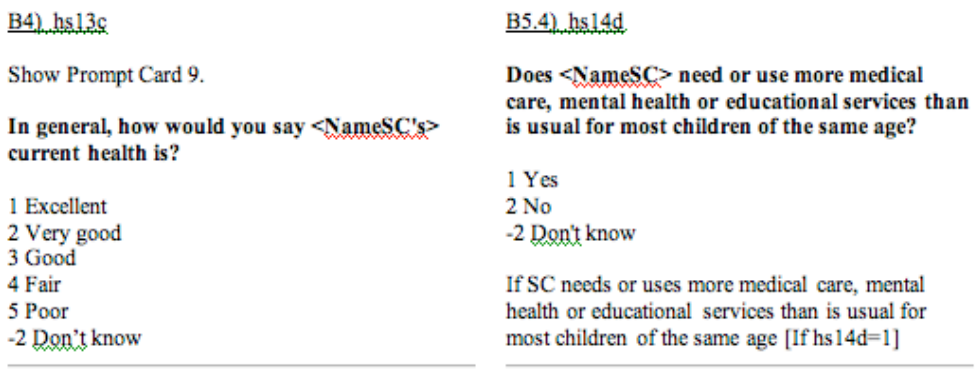


Figure 4: Example of Wave 2 interview specification

7.2 Frequencies

The frequencies are a listing of the response categories for each question and the number of cases in each category. Figure 5 provides an example of the listing.

0/1 - P1 - F2F C1.1 - Main activity - FT work				
apw01a1	**Frequency**	**Percentage (%)**	**Cumulative frequency**	**Cumulative percentage (%)**
-4	1.636675	0.03	1.636675	0.03
No	4,763.971	93.28	4,765.608	93.32
Yes	341.3922	6.68	5,107	100.00

Figure 5: Example of the weighted frequencies

The frequencies are useful for simple queries related to particular questions (e.g., how many of the births had a normal delivery, or what are the codes used for Wave 1 question A15). Variables for which there were a wide variety of responses, meaning unaltered frequencies would run for several pages (e.g., study child weight), have been rounded off to enable the grouping of responses.

7.3 Data dictionary

This is available as both an "online" version and in Excel. Both versions of the data dictionary are searchable and can be sorted. Each record describes a single variable and has the following fields:

- variable name
- variable name without age (useful for sorting)
- topic number (allows derived items to be sorted in with the input variables they come from)
- question id (i.e. variable name without age or subject/informant, useful for sorting)
- file (each of the main datasets are allocated a file name that denotes the cohort and age of the study child at each wave (I.e., Wave 1 = files B0 & K4, Wave 2 = files B2 & K6, Wave 3 = files B4 & K8 etc..).
- position in file order (the order of the variables in the files)
- wave
- cohort
- position of question in questionnaires
- person label
- child's age
- variable label briefly describing each data item
- topic
- construct
- measure
- question as found in the survey instruments
- response categories
- population with data
- SAS format
- notes field indicating other information users should know about the data item.

7.3.1 Excel data dictionary

The Excel data dictionary contains two spreadsheets, one with the complete detailed listing of variable attributes, another with a shorter listing in a print-ready format. The print-ready format contains the variable name, question, responses and population fields; however, it is not a difficult task for users to make their own printable versions if they prefer other fields.

The Excel version can be easily filtered using the drop-down menus in the first row of the spreadsheet. For example, to find all of the items on teacher practices in the lsacgr6 file (K cohort at Wave 2) first click on the drop-down menu in the “File” field as shown in Figure 6 and select “B2”. Next, repeat the process for the “Topic” field selecting “Teaching practices”.

After the search is finished all variables can be displayed by either clicking the “show all” option in each of the fields that have been filtered (see Figure 6) or by selecting “Data > Filter > Show All” from the menus.

More advanced searches can be performed using the “Custom Filter” option which produces a dialogue box to assist with your searching. For example, to find all the questions that contain the word “Internet’, go to the “question” column and open up the filter menu and click on “Custom filter”, in the dialogue box change “equals” to “contains” and type “Internet” next to this.

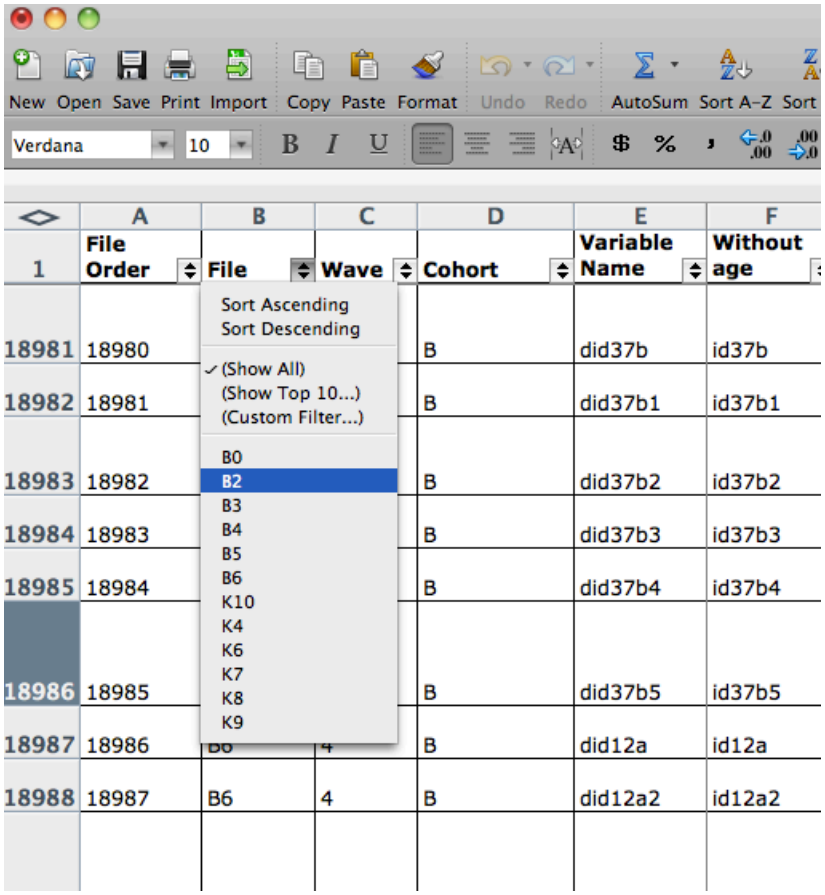


Figure 6: Example of filtering in Excel

7.3.2 Using wildcards for filtering

A good understanding of the variable naming convention is valuable for using the data dictionary. Both the online and Excel data dictionary can be searched and filtered using wildcards, which can be used to return thematically linked sets of variables. Two wildcard characters are used by both these programs:

* represents any combination of letters and characters

? represents any single character

Some examples of the use of these characters are as follows:

apw23a* returns a range of variables apw23a1a through to apw23a4b.

apw23a4? returns two variables apw23a4a and apw23a4b.

?pw23a4a shows if this variable exists over different waves.

apw23?4a shows if this variable exists for different people in the same wave.

?pw23?4a shows if this variable exists for different people in different waves.

7.3.3 Some useful tips for navigating the data dictionary

- Only items currently on the main datasets are included in the data dictionary.[7] The User Guide provides information on the composition of other datasets.
- Items on the data dictionary are in the same order as on the data files, but can easily be sorted into other orders; for example, grouping topics.
- Searching the online data dictionary finds whole words (e.g., searching for “child” won’t find “children” as well). However, an asterisk will represent any combination of characters. So searching for “child*” will find “child”, “children”, “childcare”, etc.
- The introduction page for the data dictionary contains a list of topics and constructs that can be used for finding the information you want.
- The “Question ID” field gives the variable name without any wave or person indicators. Filtering by this field is the best way to tell which questions were asked of or about which people at which wave.
- The “Topic ID” field gives the topic and associated two-digit question number for each item where this is appropriate. It can be used to link derived items with their associated input items.

Please contact the LSAC Data Management team if you need any help with using the data dictionaries.

[7] The data dictionary reflects the variables that are included in the main datasets (i.e. lsacgrb0, lsacgrb2, lsacgrb4, lsacgrb6, lsacgrb8, lsacgrb10, lsacgrk4, lsacgrk6, lsacgrk8, lsacgrk10, lsacgrk12, lsacgrk14). Items from the study child household and the PLE household modules, the NAPLAN items and the Medicare items are not in the data dictionary.

8 Data transformations

The data from many of the responses to questions have been transformed to assist data users.

8.1 Transformations to ensure consistency

LSAC contains a number of items that have been asked slightly differently across waves. Where this is logically supportable, items are recoded to match the variables produced from other waves. These recoded versions are provided in addition to the original item response. Some examples of this:

- Income is generally collected as a continuous variable; however, for the PLE in Wave 2 it was collected using five categories. To assist users in comparing the responses of different informants, an additional variable containing the continuous income information recoded into these five categories is added wherever income has been collected continuously.
- In Wave 1, respondents were asked if the child received any regular childcare from a grandparent. In Wave 2, respondents were given the option of this being a maternal or paternal grandparent. In addition to the two variables giving this information separately for maternal and paternal grandparents, an extra variable has been added for whether the child is being cared for by a grandparent.

8.2 Transformations to update information

From Wave 2 onwards, there are a number of places in the questionnaire where respondents are asked about what has happened with something since the last interview (or in the last 2 years if the study child is living in a new household). For example, in Wave 1, Parent 1 was asked how many homes the study child had lived in since birth, while in subsequent waves Parent 1 was asked how many homes the study child had lived in since the last interview. The datasets for the subsequent waves contain variables on the number of homes since the last interview and a tally of all the homes the study child has ever lived in.

8.3 Summary measures for scales

The appropriate summary measure for each scale is included, based on advice from the Consortium Advisory Group. Where it is possible to logically implement either a mean or a sum score for a psychological scale or subscale, the preference of the Consortium Advisory Group was to provide the calculation of means, except in cases where convention would dictate another scoring system. This enabled the calculation of scale level derivations where data measuring a construct has multiple contributing data items and where some contributing items are missing. Using a sum calculation for these scales would have led to the exclusion of cases with any missing data. All contributing data items to these scales are included on the datasets.

For scales where there are different sets of items for children at different ages or for different informants, multiple versions of the same scale are calculated based on just those items shared between two versions of the scale. For example, the parenting hostility scale began as a five-item measure for 0-1 year olds, but had one item dropped for children aged 4-7 years, and a further item dropped for children aged 8-9 years. On the file for 0-1 year olds three different versions of the scale are calculated: one using all five items, another using just the four items included for children aged 4-7 years, and another using just those 3 items used for children aged 8-9 years. As a general rule data users should select the variable containing the greatest number of contributing items that is appropriate for their purpose. So, for analyses using just the hostility scale at aged 0-1 years, or for those comparing the hostility scale at ages 0-1 and 2-3 years, analysts should use the five-item version. Analysts comparing hostility between the ages of 0 and 7 years should use the four-item version, and analysts comparing hostility between the ages 0 to 9 years should use the three-item version.

8.4 Outcome Index measures

A unique component of the derivation and analysis work was the development and derivation of the LSAC Outcome Index; a composite measure that indicates how children are developing. LSAC tracks the development of

children across multiple domains, and the Outcome Index provides a means of summarising this complex information for policy-makers, the media and the general public, as well as data users.

In contrast to some other indices, which focus on problems or negative outcomes, the LSAC Outcome Index, wherever possible, incorporates both positive and negative outcomes, reflecting the fact that most children have good developmental outcomes. Thus, the Outcome Index has the ability to distinguish groups of children developing poorly from those developing satisfactorily.

The rationale and methodology used to develop the Outcome Index are described in the LSAC Technical Paper No. 2,*Summarising children's wellbeing: the LSAC Outcome Index*. Papers on the derivation of the Waves 2 and 3 Outcome Index are forthcoming. Any users planning to use the Outcome Index are strongly advised to read the technical papers (available from <www.aifs.gov.au/growingup/pubs/technical/index.html>) as they contain important information about the correct use of the variable. From Wave 4 the Outcome Index is not calculated.

When undertaking longitudinal analysis involving the Outcome Index, analysts should be cautious about using outcome indices from different waves in a pooled data file as different measures may have been used at different waves to create the sub-domains.

9 Confidentialisation

Two types of data are available to data users:

- in-confidence data
- general release data.

9.1 In-confidence data

The only information not included is name, address and other contact details for the child, family, childcare agency and teacher or carer. Access to the in-confidence datasets may be granted where data users are able to demonstrate a genuine need for the additional data and that they meet the necessary additional security requirements.

9.2 General release data

In addition to the information removed for the in-confidence file, some other items have also been removed, and some items have either been transformed, had response categories collapsed or have been top-coded (i.e. recoding outlying values to a less extreme value).

The following items are removed:

- qualitative data provided by respondents
- census and postcode data for the location of carers and schools.

The following items are transformed:

- postcode—postcodes are given an indicator so that all children selected in the same postcode can be identified
- date left hospital after birth—number of days between birth and departure.

The following items have response categories collapsed (i.e. response categories combined to form an aggregate category):

- parents' occupation—output at two-digit Australian and New Zealand Standard Classification of Occupations (ANZSCO) level, or rounded off to the nearest five if ANU four ratings of occupational prestige
- occupation in previous job—output at two-digit ANZSCO level
- Socio-Economic Index for Areas (SEIFA) variables—rounded to the nearest 10
- country of birth (coded as 0 if fewer than five contributors)
- religion (coded as 0 if fewer than five contributors)
- language other than English (LOTE) (coded as 0 if fewer than five respondents).

The following data items are top-coded:

- income
- housing costs
- child support paid by Parent 2
- children and parents' current height, weight and waist circumference
- number of hours spent in child care.

10 Data imputation

Limited imputation of data is undertaken in LSAC. In general, imputation occurs only when there is clear contradiction between data items and there is a good reason to believe one item over the other. Some basic principles are applied for this task.

10.1 Virtual roll-forward

“Roll-forward” is the term in CAI design that refers to the use of data from a previous wave of data collection to determine the questions that need to be asked in a subsequent wave. For Wave 2 a limited set of data was rolled forward, largely to assist with the household composition module. Time and resource implications meant that roll-forward could not be used in some other parts of the questionnaire where it may have reduced respondent burden.

For example, in Wave 2, respondents were again asked about the age the child stopped being breastfed, in order to obtain the information from those cases where this had not yet happened at the time of Wave 1. In re-asking this question, some respondents gave different answers to their Wave 1 responses. Given that recollection of respondents is likely to be more accurate closer to the event (i.e. the cessation of breastfeeding), it was decided that in cases where Wave 1 data exists, the Wave 1 value is taken as correct and the Wave 2 value is ignored (i.e. as if the Wave 1 data had been rolled forward and the question never asked in Wave 2). This means a single variable is produced that represents the best estimate from the two waves of data. (Users are able to tell at which wave the timing data was collected by referring to the question from each wave asking if the child is still being breastfed.)

Note: From Wave 3 onwards there is a greater use of roll-forward, which reduced the number of situations where such conflicts could occur.

10.2 Longitudinal contradictions

Another possible contradiction in the data may occur where respondents report at a subsequent wave that an event occurred at a time before a previous wave, when the previous wave’s data indicated that this event hadn't happened yet.

In these cases the time of the previous wave is treated as the time of the event. For example, if a parent reported at Wave 2 that the child stopped being breastfed after two months; however, at Wave 1 the child was three months old and was reported as still being breastfed, the age of breastfeeding cessation would be set to three months.

This strategy for fixing the time of an event is also used for the:

- date when new members joined the household
- length of attendance at a particular child care facility
- date left the household for Wave 1 members and temporary members (bf14m1, bf14m2, etc.)
- age stopped breastfeeding (zf05c)
- age first had non-breast milk (zhb07)
- age first had solid food (zhb10)
- age entered child care arrangements (bpc11a, bpc11b, etc.)
- age last lived with two biological parents (bpe23c).

10.3 Other imputations

On inspection of the data, problems were revealed in a small number of items. These problems were solved using imputation and are listed below.

- Employment status: Some assumptions are made to assist in coding the parent to employed, unemployed or not in the labour force where missing values were present.

- Type of educational program (K cohort Wave 1): There appeared to be some confusion with parents and interviewers as to whether the child was in pre-school or pre-Year 1 at school. The type of education program variable was amended based on the teacher data and other information provided in the questionnaire.
- Parental income: Outlying values, particularly those with responses to other questions (e.g., categorical income, sources of income) that make the income value appear incorrect, were adjusted. For further information about imputations related to parental income, see LSAC Technical Paper No. 14 (available from <www.growingupinaustralia.gov.au/pubs/technical/index.html>).
- Parental height: It was found that there were some changes in height between waves for some parents of study children. While most were minor (most likely due to estimation error) some were more substantial, and called into question the reliability of differences in body mass index recordings between waves.
- Time use diary data: Responses were recorded by marking an oval to indicate whether an activity/situation occurred in each 15-minute time period. A number of "false positives" were discovered in the Wave 1 TUD data. Imputation was used to reduce the number of false positives. A number of imputations were also performed to improve data quality in all three waves.

Further details of these imputations are given in *Data Issues Waves 1 to 6* (available on the study website<www.growingupinaustralia.gov.au/pubs/issues/index.html>)

11 Survey methodology

LSAC employs a cross-sequential design that follows two cohorts of children, initially aged 0-1 years (B cohort) and 4-5 year olds (K cohort) in 2004.

Families are visited by interviewers every two years to collect data for the main waves of the study. In the "between" years, a mail-out survey was conducted at Waves 1.5, 2.5 and 3.5 to help maintain contact with families and obtain some additional information. At Waves 4.5 and 5.5, a web form was used primarily to update contact details.

The key features of the initial sample design and methodology for each wave are included in this section. A full description of the sample design is given in LSAC Technical Paper No. 1, and details of the weighting and non-response analysis are given in subsequent technical papers.

(For technical papers see <www.aifs.gov.au/growingup/pubs/technical/index.html>).

11.1 Sample design

A two-stage clustered sample design was employed, first selecting postcodes then children, with the clustered design allowing analysis of children within communities and producing cost savings for interviews.

Stratification was used to ensure proportional geographic representation for states/territories and capital city statistical division/rest of state areas. The sample was stratified by state, capital city statistical division/balance of state and two strata based on the size of the target population in the postcode.

Postcodes were selected with probability proportional to size selection where possible, and with equal probability for small population postcodes. Children from both cohorts were selected from the same 311 postcodes. Some remote postcodes were excluded from the design, and the population estimates were adjusted accordingly.

Children were selected with approximately equal chance of selection for each child (about one in 25).

Apart from some remote areas, the sample was selected to be representative of all Australian children (citizens and permanent residents) in each of two selected age cohorts:

- children born March 2003-February 2004 (B cohort)
- children born March 1999-February 2000 (K cohort).

11.1.1 Sample selection and recruitment

The sample was selected from Medicare Australia's enrolment database. Within the selected postcodes, the population was ordered by date of birth and then a random start and skip applied to select the children. The actual number of children selected depended on which stratum the postcode was in but for most postcodes the aim was to recruit about 20 children per cohort per postcode.

The selection of children and corresponding Wave 1 fieldwork occurred in 4 phases, partly to reduce the age range of children at interview and partly because some of the target population had not been born at the time of the first phase selection.

Families of 18,800 selected children received letters of invitation to take part in the study sent by Medicare Australia. Families could "opt-out" of the study by phoning a 1800 number or returning a reply-paid slip. Medicare Australia 1800 staff were given training about the study and were able to answer queries and make notes of other information (for example, telephone numbers).

After a 4-week opt-out period, Medicare Australia gave the contact names and addresses of remaining families to I-view, the Wave 1 data collection agency. I-view then sent another letter to families saying when an interviewer would be in their area.

I-view maintained a 1800 number for families selected in the study, which was transferred to the ABS who took responsibility for the data collection from Wave 2 onwards.

11.2 Development and testing of survey instruments

11.2.1 Pre-testing

Pre-testing of new material and processes is undertaken at each wave of the study, comprising small-scale pre-tests and cognitive interviews. In Waves 1 and 2, more formal piloting was also undertaken. Small-scale testing is also undertaken for the between-wave surveys.

Wave 1

- Development began in March 2002.
- Small-scale pre-testing occurred in September-October 2002.
- A pilot test with about 50 families from each cohort was conducted in March-April 2003.

Wave 2

- Development began in July 2004.
- Small-scale pre-testing occurred in September-October 2004.
- A pilot test with 86 families was conducted in April 2005.

Wave 3

- Development began in March 2006.
- Pre-testing occurred in a number of stages from mid 2006 to March 2007.
- No pilot test was required.

Wave 4

- Development began in February 2008.
- Pre-testing occurred in a number of stages from mid August 2008 to June 2009.
- No pilot test was required.

Wave 5

- Development began in February 2010.
- Pre-testing occurred in a number of stages from mid June 2009 to March 2010.
- No pilot test was required.

Wave 6

- Development began in May 2012.
- Pre-testing occurred in a number of stages from August 2012 to September 2013.
- No pilot test was required.

11.2.2 Dress rehearsal

In Wave 1, a dress rehearsal (DR) sample of 526 families was recruited to test the content and processes intended for the main waves of the study. Over 1000 children were initially selected from 25 postcodes in Victoria, Sydney and rural/remote New South Wales and Queensland. Postcodes in Victoria were selected at random, but the other postcodes were selected as areas that may provide challenges to the data collection process. Other dress rehearsals have also been completed.

- Wave 1 DR—August-November 2003 (526 families interviewed)
- Wave 2 DR—September-November 2005 (423 families interviewed)
- Wave 3 DR—July-October 2007 (420 families interviewed)
- Wave 4 DR—July-October 2009 (387 families interviewed)

- Wave 5 DR—July-August 2011 (451 families interviewed)
- Wave 6 DR—June-August 2013 (351 families interviewed)

After each dress rehearsal, both processes and content have been refined to increase efficiency and reduce the time in the home.

11.3 Data collection

11.3.1 Interview length

In Wave 1, an average of 126 minutes was allowed for time in the home by the interviewer. In-home data collection with the B cohort averaged about 1.5 hours, while interviews for the K cohort averaged about 2.5 hours.

In Wave 2, although an average of 90 minutes had been allowed for the time in the home, the actual time was shorter, averaging 66 minutes for the B cohort and 85 minutes for the K cohort.

In Wave 3, an average of 100 minutes was allowed for time in the home; the actual time was 91 minutes for the B cohort and 98 minutes for the K cohort.

In Wave 4, an average of 110 minutes was allowed for time in the home; the actual time was 102 minutes for the B cohort and 108 minutes for the K cohort.

In Wave 5, an average of 110 minutes was allowed for time in the home; the actual time was 98 minutes for both cohorts.

In Wave 6, and average of 110 minutes was allowed for time in the home; the actual time was 108 minutes for the B cohort and 116 minutes for the K cohort.

11.3.2 Interviewers

As part of a standard ABS interviewer induction, ABS interviewers receive two weeks of intensive training across a range of standard procedures and practices. All interviewers received 8 hours of home learning (this included a computer-based learning module, home study exercises and reading of interviewer instructions).

In Wave 1, 150 interviewers and field supervisors from I-view were trained during a series of 4-day sequential training courses conducted in Melbourne, Brisbane, Perth and Sydney during February and early March 2004. The principal trainers were the same for all courses, ensuring consistency in training.

Psychologists conducted the training for “Who am I?”, the PPVT and the interviewer observations. A large part of the training involved practice interviews, with one day devoted to interviews with parents and children.

For Wave 2, 147 interviewers from ABS were trained in a series of 3-day training courses in Sydney, Melbourne, Brisbane and Perth during March and April 2006. Two training teams were used, comprising staff from both AIFS and ABS. This time, AIFS staff undertook the direct assessment training, after receiving training from a child psychologist (the use of computer-assisted interviewing for the direct assessments helped ensure the consistent administration of these assessments).

For Wave 3, 176 interviewers from ABS were trained in a series of 2-day training courses in Brisbane, Melbourne, Sydney and Perth during March and April 2008. Interviewers who had not worked on LSAC previously were given background training in LSAC before the 2-day course commenced. Two training teams were used, comprising staff from ABS, AIFS and DSS. Again, AIFS staff undertook the direct assessment training.

For Wave 4, 181 interviewers from ABS were trained in a series of 3-day training courses in Brisbane, Melbourne, Sydney and Perth. Two training teams were used, comprising staff from the ABS, AIFS and DSS. As in previous waves, AIFS staff undertook the direct assessment training.

For Wave 5, 198 interviewers from ABS were trained in a series of 3-day training courses in Brisbane, Melbourne, Sydney, Adelaide and Perth. New-to-LSAC interviewers (defined as anyone who did not participate in Main Wave 4) attended the first day of classroom training where topics such as “Background to the study”, “Physical measurements”, “Direct assessments” and “Notebook security” were covered. All interviewers attended days 2 and

3 when the P1 interviews and the K and B child interviews were covered in detail (apart from what was done on day 1). New interviewers were teamed with an experienced interviewer, allowing for mentoring throughout the training course and for the new interviewers to be the interviewer during practice sessions.

For Wave 6, 200 interviewers from ABS were trained in a series of 4-day training courses in Brisbane, Melbourne, Sydney, Adelaide and Perth. All interviewers attended the full 4-day training program due to the large amount of new content and procedures. During the practice sessions, interviewers were split into groups of three (rather than pairs as in previous waves). This allowed for a more realistic practice with each interviewer taking the role of the parent, child, and interviewer. Where possible in the training sessions and in the practice sessions, new LSAC interviewers were paired with experienced LSAC interviewers. ABS staff conducted all the training.

11.3.3 Fieldwork periods

Wave 1

Selected postcodes were divided into two groups for maximum field efficiency. The target population was also divided into two groups: children born March-August (older) in one group and children born September-February (younger) in the other.

The fieldwork was divided into four phases:

- Phase 1 started in mid March 2004 for the older children in the first group of postcodes
- Phase 2 started at the end of April for the older children in the second group of postcodes
- Phase 3 started in June for the younger children in the first group of postcodes
- Phase 4 started in late July for the younger children in the second group of postcodes.

Follow-up continued throughout 2004. The blue line in Figure 6 shows the distribution of interviews over time for Wave 1 fieldwork.

Wave 2

Again there were broadly 4 fieldwork periods, although the dates for these varied from state to state. Regional offices of the ABS were able to organise the work to suit the availability of interviewers and other work. As far as possible, ABS tried to interview the children born in March-August in the first 2 periods, and children born in September-February in the later fieldwork periods. 84% of the interviews were conducted prior to September 2006.

Figure 6 shows the distribution of interviews over time for Wave 2 fieldwork. Fieldwork started later than in Wave 1 due to the additional work required to prepare the CAI instrument.

Wave 3

Fieldwork was organised as per Wave 2. The green line in Figure 7 (below) shows the distribution of interviews over time for Wave 3 fieldwork.

Wave 4

As the children get older, the age differences within a cohort are less significant. To assist the efficiency of work allocations to interviewers, the focus in Wave 4 was more on the location of the sample and interviewers with less emphasis given to following interviews within the set phases. The dark blue line in Figure 7 shows the distribution of interviews over time for Wave 4 fieldwork.

Wave 5

Fieldwork was organised based on the location of the sample and interviewers..Figure 7 shows that the distribution of interviews for Wave 5 fieldwork was more spread out across the months than for previous waves.

Wave 6

Fieldwork was organised based on the location of the sample and interviewers. Figure 7 shows that the distribution of interviews for Wave 6 fieldwork was distributed across the months, similar to Wave 5.

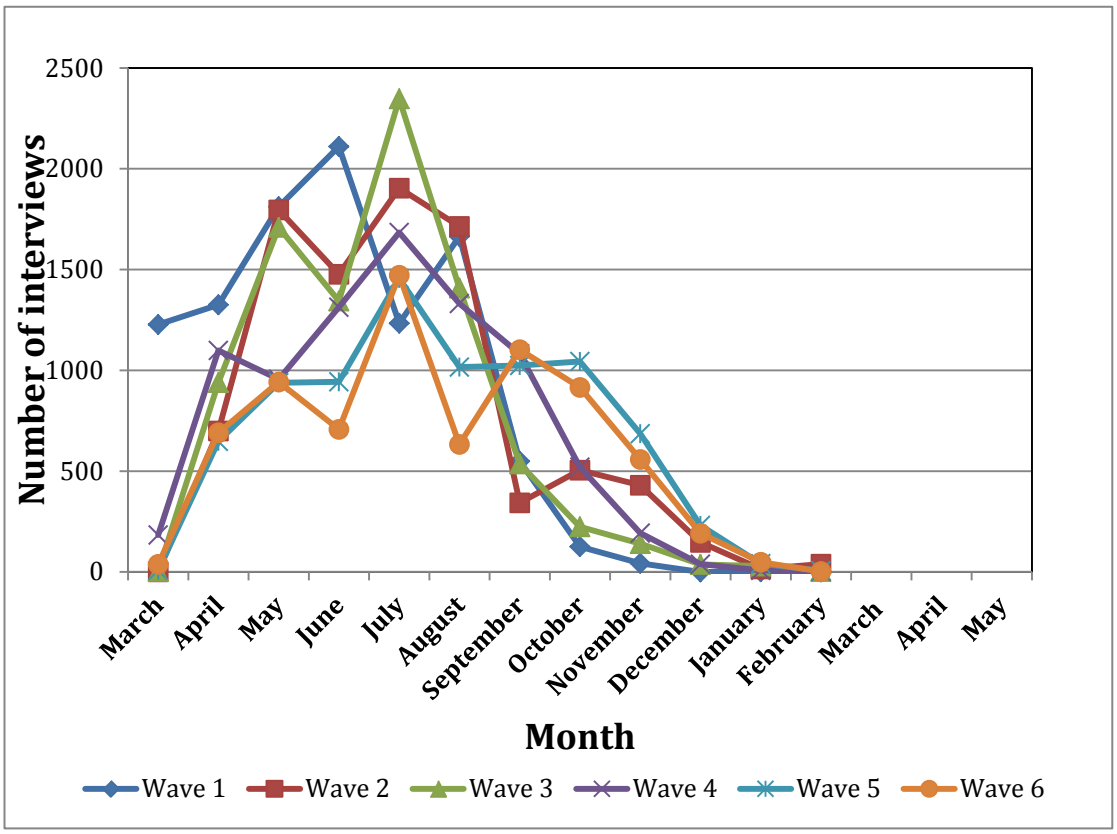


Figure 7: Month of interview for study families in Waves 1 to 6

11.3.4 Contact process

Wave 1

For most families, the interviewer only had the name and address of the Medicare cardholder and which cohort the child was in. In a small number of cases, families who were keen to participate had contacted the 1800 numbers and supplied phone numbers and/or best times to call.

Interviewers were required to make up to six visits to the address, at different times of the day and on different days of the week. A major challenge was that 7% of addresses were post office box addresses, and although families with these addresses were specifically requested to make contact with the 1800 number to supply a residential address, only a small proportion did so. In addition, many of the residential addresses held by Medicare were found to be out of date by the time the interviewers visited. Interviewers made significant attempts to locate families for whom they did not have a current residential address, by referencing the White Pages and electoral rolls and speaking with neighbours and other local contacts.

Between waves

Contact is maintained with study families between waves by sending birthday cards, annual calendars and newsletters and via the between-wave mail-out and online questionnaires. These processes have resulted in some families contacting the ABS to update their contact information, which helps when trying to arrange appointments for the main waves of interviewing.

Subsequent waves

Pre-interview letters plus a brochure outlining the processes for that wave were sent to all families who had not opted out of the study since the previous wave, unless it was confirmed that the address was out of date. Interviewers then followed up with a telephone call to make an appointment for an interview. If the contact information was out of date, the interviewers tried to contact secondary contacts of Parent 1 (these details were given by Parent 1 in Wave 1 and are updated each wave) to locate the family. One visit to the address was also made. If the family could not be located, the interviewer referred this back to the office for tracking.

After an appointment for interview was made, the interviewer confirmed the appointment the day before the appointment.

11.3.5 Foreign language interviews

Wave 1

As part of the Medicare Australia mail-out, a brochure was included with information about the study in nine languages. Medicare Australia staff made use of the Telephone Interpreter Service (TIS) to assist with calls where required.

Apart from this brochure, no other study material was (or has been) translated into other languages, and instead interpreters were used. An interpreter was required in 3% of interviews, with over 50 languages involved. In most cases (138), a member of the family or friend was preferred as the interpreter. In 76 cases, an I-view employee was able to act as interpreter, and in 96 cases, an interpreter was employed.

Wave 2

A total of 110 interviews (1%) were conducted in a language other than English, in 23 different languages. Family or friends assisted in 58 cases, ABS interpreters helped in 37 cases, and a TIS interviewer was used for 15 families. An interpreter was arranged whenever requested or judged necessary by the interviewer. The reduction in use of interpreters between waves is presumably due to the increased confidence in English that had been gained by respondents in this time.

Wave 3

A total of 97 interviews needed an interpreter, in 24 languages. Family or friends assisted in 58 cases, ABS interpreters helped in 31 cases, and a TIS interviewer was used for 8 families.

Wave 4

A total of 93 interviews needed an interpreter, in 26 languages. Family or friends assisted in 50 cases, ABS interpreters helped in 29 cases, and a TIS interviewer was used for 14 families.

Wave 5

A total of 81 interviews needed an interpreter, in 18 languages. Family or friends assisted in 47 cases, ABS interpreters helped in 24 cases, and a TIS interviewer was used for 10 families.

Wave 6

A total of 64 interviews needed an interpreter, in 17 languages. Family or friends assisted in 42 cases, ABS interpreters helped in 18 cases, and a TIS interviewer was used for 4 families.

11.3.6 Indigenous communities

Although the sample selection process excluded 40% of areas classified as remote by the ABS (areas that typically have a high Indigenous population) there were still a number of postcodes selected that contained some remote Indigenous communities, hence strategies have been put in place to enumerate these communities.

Where feasible, communities were visited or telephoned, and personal contact made with a number of community organisations from whom assistance was gained to identify whether families were in residence and willing to be interviewed. Travel to remote communities was only undertaken if there was an appointment for an interview.

Aboriginal and Torres Strait Islander families are included in representative numbers in non-remote centres. However, there has been a higher rate of attrition from the study among these families. See the weighting and non-response technical papers for more details (available from <www.aifs.gov.au/growingup/pubs/technical/index.html>).

11.3.7 Remote areas

In the initial sample there were 12 postcodes selected in areas classified as "remote" by the ABS Australian Standard Geographic Classification (ASGC) Remoteness Classification. Interviewers were either recruited from these areas or travelled to these areas when the field agency did not have a suitable interviewer in the locality.

Where visits were not possible, telephone interviews were conducted: 12 (0.12%) in Wave 1, 42 (0.46%) in Wave 2, 87 (0.10%) in Wave 3, 83 (0.99%) in Wave 4, 73 (0.91%) in Wave 5 and 59 (0.81%) in Wave 6. The increasing number is due to sample dispersion.

11.4 Fieldwork response

11.4.1 Wave 1 recruitment

The final response to the recruitment of children was 54% of those families who were sent the initial letter by Medicare Australia. The response rate was higher for the B cohort, with 57% of families (5,107) agreeing to take part, compared with 50% of K cohort families (4,983).

About 35% of families who were sent the initial letter refused to take part in the study. The main reasons given to interviewers for not participating in the study were: not interested/too busy (57%), not capable/moving/overseas (9%), husband refused (9%), and illness/death (8%). The remaining 13% of families could not be contacted, despite intensive efforts from interviewers.

Non-response analysis was undertaken to determine how representative the sample is of all Australian children in the scope of this study, and adjustments have been made to the survey weights to allow for this. For further information on the weighting and non-response, see LSAC Technical Paper No. 3, *Wave 1 weighting and non-response analysis* (available from <www.aifs.gov.au/growingup/pubs/technical/index.html>).

Response in later waves

Table 9 summarises the response from families in later waves, using the Wave 1 sample and "available" sample as the bases for comparisons.

Table 9: Sample size and response rate for each wave and cohort of LSAC

	B cohort			K cohort			Total		
	N	Resp. rate of Wave 1 (%)	Resp. rate of available sample (%)	N	Resp. rate of Wave 1 (%)	Resp. rate of available sample (%)	N	Resp. rate of Wave 1 (%)	Resp. rate of available sample (%)
Main waves									
Wave 1 original	5,107	100		4,983	100		10,090	100	
Wave 2 available[1]	5,047	98.8		4,913	98.6		9,960	98.7	
Wave 2	4,606	90.2	91.2	4,464	89.6	90.9	9,070	89.9	91.1
Wave 3 available	4,971	97.3		4,829	96.9		9,800	97.1	
Wave 3	4,386	85.9	88.2	4,332	86.9	89.7	8,718	86.4	89.0
Wave 4 available	4,929	96.5		4,774	95.8		9,703	96.2	
Wave 4	4,241	83.0	86.0	4,164	83.5	87.2	8,405	83.3	86.6
Wave 5 available	4,884	96.6		4,735	95.0		9,619	95.3	
Wave 5	4,085	80.0	91.1	3,956	79.4	83.5	8,041	79.7	83.6
Wave 6 available	4,483	87.8		4,395	88.2		8,878	88.0	
Wave 6	3,764	73.7	84.0	3,537	71.0	80.5	7,301	72.4	82.2
Between-waves									
Wave 1.5 sent	5,061	99.1		4,935	99.0		9,996	99.1	
Wave 1.5 returned	3,573	70.0	70.6	3,584	71.9	72.6	7,157	71.0	71.6
Wave 2.5 sent	4,859	95.1		4,712	94.6		9,571	94.9	
Wave 2.5 returned	3,268	63.5	64.0	3,287	65.5	66.0	6,555	63.4	65.0
Wave 3.5 sent	4,772	93.4		4,641	93.1		9,413	93.3	
Wave 3.5 returned	3,012	59.0	63.1	2,972	59.6	64.0	5,984	59.3	63.6

Notes: Excludes in-between Waves 4.5 and 5.5 where the data is not relevant for users of the LSAC datasets. They were used only to update contact details. [1]Available sample excludes those who opted out of the study between waves. Some additional families also opted out permanently during the fieldwork process. [2] Those who had a home visit.

Table 10 details the reasons why interviews were not obtained in Waves 2 to 6.

Table 10: Response status and reasons for non-response by wave

	Wave 2		Wave 3		Wave 4		Wave 5		Wave 6	
Response status	N	%	N	%	N	%	N	%	N	%
Responding	9,070	91.1	8,718	89.0	8,405	86.6	8,041	83.6	7,301	82.2
Refusal	284	2.8	436	4.4	637	6.6	774	8.0	938	10.6
Non-contact	540	5.4	552	5.6	526	5.4	715	7.4	555	6.3
Away entire enumeration period	61	0.6	93	1.0	135	1.4	88	0.9	39	.4
Death of study child	5	0.1	1	0.01	0	0	1	0.01	3	0.0
Other	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	43	0.5
Total starting sample	9,960	100.0	9,800	100.0	9,703	100.0	9,619	100.0	8,879	100

12 Important issues for data analysis

A data issues paper for Waves 1 to 6 is available and details issues that have been identified over the course of the study. Data users should be aware of these issues when using LSAC data for their analysis. (See <www.growingupinaustralia.gov.au/pubs/issues/index.html>)

Other important issues are addressed below.

12.1 Weighting and external validity

The LSAC study design, based on a complex probability sample, is specifically designed to produce valid estimates at the population level. Unlike clinically based or convenience samples, the LSAC sample is population based by design. By properly accounting for the survey design when analysing the data, it is possible not only to make inferences about the children and families participating in the study but to make valid inferences about the entire population of children in the relevant age groups.

The LSAC sampling strategy has three important elements that distinguish it from a simple random sample (SRS):

- *stratification* to ensure proportional representation of all states and both capital city and ex-metropolitan areas
- *clustering* by postcode to both reduce field enumeration costs and allow the study of community-level effects on children's development and wellbeing
- *weighting* to adjust for potential non-response bias and to provide population estimates.

It is the responsibility of data users to determine when and how each of these needs to be accounted for when developing their analyses.

12.1.1 Stratification

Stratification, by state and part of state, was employed to ensure that all geographic areas within Australia are represented in the sample in proportion to their population. This produces a more even distribution of the sample across geographic areas than could be expected from a simple random sample.

The use of stratification can be expected to reduce standard errors compared with a simple random sample with no control over the geographic spread of the sample. As such, when trying to extrapolate to the population, the stratification should be incorporated in the analysis of results from the survey in order to correctly calculate standard errors and confidence intervals.

12.1.2 Clustering

The use of clustering in the sample design has important consequences for the analysis of data from the study. Clustering is useful in reducing the field costs associated with the survey enumeration. Clustering also has the added benefit of making possible the analysis of community-level effects, by ensuring that sufficient sample is selected from each postcode included in the survey.

However, the use of clustering violates the standard assumption of independence of the observations that is fundamental to many statistical routines in major statistical packages. When children or carers have more similar characteristics within a given postcode than children or carers selected purely at random, the responses within postcodes will be correlated. This correlation will lead to an increase in the standard errors and size of the confidence intervals. The extent of this increase is measured by the *design effect*, which is the ratio of the variance of an estimate from the survey to the variance that would have been achieved by a simple random sample of the same size.

Failure to account for clustering in the analysis can lead to under-estimating the size of standard errors and confidence intervals. In some circumstances this can result in misleading conclusions of statistical significance.

12.1.3 Weighting

The Wave 1 weights provided in the LSAC data files take into account both the probability of selecting each child in the study and an adjustment for non-response. An analysis of possible differences in the characteristics of respondents and non-respondents was undertaken and identified two factors associated with the probability of participating in the survey—whether the mother speaks a language other than English at home, and whether the mother has completed year 12. Both of these factors were incorporated into the Wave 1 survey weighting so that, to the best extent possible, the use of the sample weights offset the bias that may be introduced into the data due to differential non-response patterns.

At each subsequent wave of data collection, weights have been adjusted to account for the differential probability of response as estimated by regression. The weights are then calibrated back to the stratum benchmarks and a small number of cases have their weights top or bottom coded to prevent any case having too great or small an effect on the data.

From Wave 3 onwards, it was required to produce longitudinal as well as cross-sectional weights for the first time. Cross-sectional weights adjust the sample attained at current wave to be representative of the population at the time of selection (i.e., when first interviewed) while longitudinal weights do the same for the sample that has responded to all waves of the survey.

More detailed information on the weighting variables can be found in LSAC Technical Papers No. 3, 5, 6, 9, 10 and 15 (available at <www.aifs.gov.au/growingup/pubs/technical/index.html>).

Three types of weight are included in the LSAC datasets:

- *Child population weights*—these weights are used to produce population estimates based on the LSAC data (e.g., based on LSAC data there are 22,464 children born in March 2003 to February 2004 in Australia that were never breastfed).

  The sum of the responding B cohort child population weights is 243,026 and the sum of the K cohort child population weights is 253,202, which are the ABS estimated resident population counts of children aged 0 years and 4 years, respectively, at end March 2004, adjusted for the remote parts of Australia that were excluded from the study design.
- *Child sample weight*—this is the child population weight rescaled such that the sum of the weights matches the number of children in the sample (e.g., 5,107 B cohort and 4,983 K cohort in Wave 1).

  This weight is used in analyses that expect the weights to sum to the sample size rather than the population, particularly when tests of statistical significance are involved.
- *Time use data day weight* (for Waves 1, 2 and 3 only)—this is the sample weight adjusted so that each day of the week receives equal weight in analyses of time use data.

Data files for Wave 1 and Wave 2 each have one population weight and one sample weight. Given that there are no cases that responded to Wave 2 that didn't respond to Wave 1, these weights can be used for both longitudinal and cross-sectional analyses.

At Wave 3, two sample weights and two population weights are necessary as this is the first time that respondents could return to the study after missing a wave. The first of these weights the full Wave 3 sample and should be used for cross-sectional analyses. The second weights the sample that has responded to all waves, and should be used for longitudinal analyses.

A complete list of LSAC weighting variables is given in Tables 11 (B cohort) and 12 (K cohort).

Table 11. Weighting variables for B cohort

Variable name	Cohort	Type	Waves cases responded to	Used for
aweight	B	Population	1	Wave 1 cross-sectional analyses

Variable name	Cohort	Type	Waves cases responded to	Used for
aweights	B	Sample	1	Wave 1 cross-sectional analyses
aweightd	B	Day	1	Wave 1 cross-sectional analyses
bweight	B	Population	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
bweights	B	Sample	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
bweightd	B	Day	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
cweight	B	Population	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
cweights	B	Sample	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
cweightd	B	Day	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
bcwt	B	Population	1, 2 & 3	Longitudinal analyses involving all Wave up to Wave 3.
bcwts	B	Sample	1, 2 & 3	Longitudinal analyses involving all Wave up to Wave 3.
bcwtd	B	Day	1, 2 & 3	Longitudinal analyses involving all Wave up to Wave 3.
dweight	B	Population	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
dweights	B	Sample	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
eweight	B	Population	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
eweights	B	Sample	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
bdwt	B	Population	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
bdwts	B	Sample	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
cdwt	B	Population	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves 1, 3 & 4
cdwts	B	Sample	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves 1, 3 & 4
bcdwt	B	Population	1, 2, 3 & 4	Longitudinal analyses involving all Wave up to Wave 4.
bcdwts	B	Sample	1, 2, 3 & 4	Longitudinal analyses involving all Wave up to Wave 4.
bcdewt	B	Population	1, 2, 3, 4 & 5	Longitudinal analyses involving all Wave up to Wave 5.
bcdewts	B	Sample	1, 2, 3, 4 & 5	Longitudinal analyses involving all Wave up to Wave 5.
fweight	B	Population	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
fweights	B	Sample	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
bcdefwt	B	Population	1, 2, 3, 4, 5 & 6	Longitudinal analyses involving all Waves up to Wave 6.
bcdefwts	B	Sample	1, 2, 3, 4, 5 & 6	Longitudinal analyses involving all Waves up to Wave 6.

Table 12. Weighting variables for K cohort

Variable name	Cohort	Type	Waves cases responded to	Used for
cweight	K	Population	1	Wave 1 cross-sectional analyses
cweights	K	Sample	1	Wave 1 cross-sectional analyses
cweightd	K	Day	1	Wave 1 cross-sectional analyses
dweight	K	Population	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
dweights	K	Sample	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
dweightd	K	Day	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
eweight	K	Population	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
eweights	K	Sample	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
eweightd	K	Day	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
dewt	K	Population	1, 2 & 3	Longitudinal analyses involving all Wave up to Wave 3.
dewts	K	Sample	1, 2 & 3	Longitudinal analyses involving all Wave up to Wave 3.
dewtd	K	Day	1, 2 & 3	Longitudinal analyses involving all Wave up to Wave 3.
fweight	K	Population	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
fweights	K	Sample	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
dfwt	K	Population	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
dfwts	K	Sample	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
efwt	K	Population	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves1, 3 & 4
efwts	K	Sample	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves 1, 3 & 4
defwt	K	Population	1, 2, 3 & 4	Longitudinal analyses involving all Waves up to Wave 4.
defwts	K	Sample	1, 2, 3 & 4	Longitudinal analyses involving all Waves up to Wave 4.
gweight	K	Population	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
gweights	K	Sample	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
defgwt	K	Population	2, 3, 4 & 5	Longitudinal analyses involving all Waves up to Wave 5.
defgwts	K	Sample	2, 3, 4 & 5	Longitudinal analyses involving all Waves up to Wave 5.
hweight	K	Population	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
hweights	K	Sample	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
defghwts	K	Sample	1, 2, 3, 4, 5 & 6	Longitudinal analyses involving all Waves up to Wave 6.
defghwt	K	Population	1, 2, 3, 4, 5 & 6	Longitudinal analyses involving all Waves up to Wave 6.

12.1.4 Survey estimation and analysis techniques

Survey estimation and analysis techniques are available that can take all three key features of the study design into account, and many of these techniques are now included in commercially available software. Incorporating the

study design features into analyses of the study can produce externally valid results at the full population level. Estimates of means, proportions and totals incorporating the study design provide the best estimate of the true means, proportions and totals within the total population.

Analytic techniques, particularly modelling, aim at exploring relationships within the data and are able to estimate the best fitting model for the underlying population, not just the best fitting model for the sample, when properly applied to account for the study design.

12.1.5 Useful references

An overview of population survey methods is given by Levy and Lemeshow (1999). They discuss the use of stratification, weighting and clustering in survey design and the impact it has on the analysis of sample survey data.

For a thorough discussion of the mathematical techniques used to analyse data from complex surveys, see Chambers and Skinner (2003).

12.1.6 Software

There is now a range of software available from a number of vendors that supports the analysis of data from complex survey designs incorporating stratification, clustering and weighting. These include SAS (using the SURVEYMEANS and SURVEYREG procedures), STATA (using the svy commands), and SPSS (through the SPSS Complex Samples add-on module), as well as software packages specifically designed for the analysis of sample survey data such as WesVar and SUDAAN.

Use of the appropriate analytic techniques from one or more of these packages is recommended for researchers analysing the LSAC data. Results that properly account for the sample design features will have the greatest external validity and should be appropriate for drawing inferences about the total population of children from which the sample was taken.

Figure 8 shows a template for using the SURVEYREG and SURVEYMEANS procedures in SAS.

```
proc surveyreg data=<filename> total=<stratumfile>;
   stratum stratum;
   cluster pcodes;
   model <standard SAS model details>;
   weight weights;
run;

proc surveymeans data=<filename> total=<stratumfile>;
   stratum stratum;
   cluster pcodes;
   var <variable names>;
   weight weights;
run;

Where:
stratum: is a variable you can calculate for lsac0 using the formula:
   stratum=int(mod(hicid,10000000)/100000);

pcodes: is the postcode of selection (already on the data file)

weights: is the sample weight (preferred to the population weight for this analysis)

<stratumfile> is a file that contains the number of Primary Sampling Units (in this case postcode clusters) in each stratum.  It is included on the data CD or can be set up using the following code.

data stratum;
 input stratum _total_; datalines;
11 295
13 168
14 160
```

```
21 202
22 58
23 95
24 316
31 116
33 121
34 108
41 110
43 34
44 131
51 82
52 86
53 32
54 103
61 28
63 38
71 9
73 3
74 1
81 23
;
run;
```

Figure 8: SURVEYREG and SURVEYMEANS procedures in SAS

12.2 Unit of analysis

The child is the unit of selection in LSAC and estimates produced from this survey are of children, not of parents or families. It is important that this point is understood when producing population estimates from this survey.

Using the estimates to count families/parents will produce an over-count of the number of families/parents, due to the multiple (or over) counting of children from multiple births. Although this will not make a huge difference to the actual numbers, it may be important in the interpretation of the information and in comparing data from other sources.

Although it is possible to produce “family” weights, it is not considered a worthwhile use of resources given the small number of analyses this could possibly meaningfully affect.

12.3 Age at interview

Different ages of children should be accounted for in any analyses focused on age-dependent measures such as cognitive and motor development. Figures 9 and 10 show the age distribution of the two cohorts at each wave. The figures show the age of the child as a base figure (i.e. 0, 2, 4, 6, 8, 10, 12 and 14 years) plus a number of months. For example, a B-cohort child aged 3 years and 1 month at time of interview in Wave 2 is shown against “13” on the x-axis (see the red line).

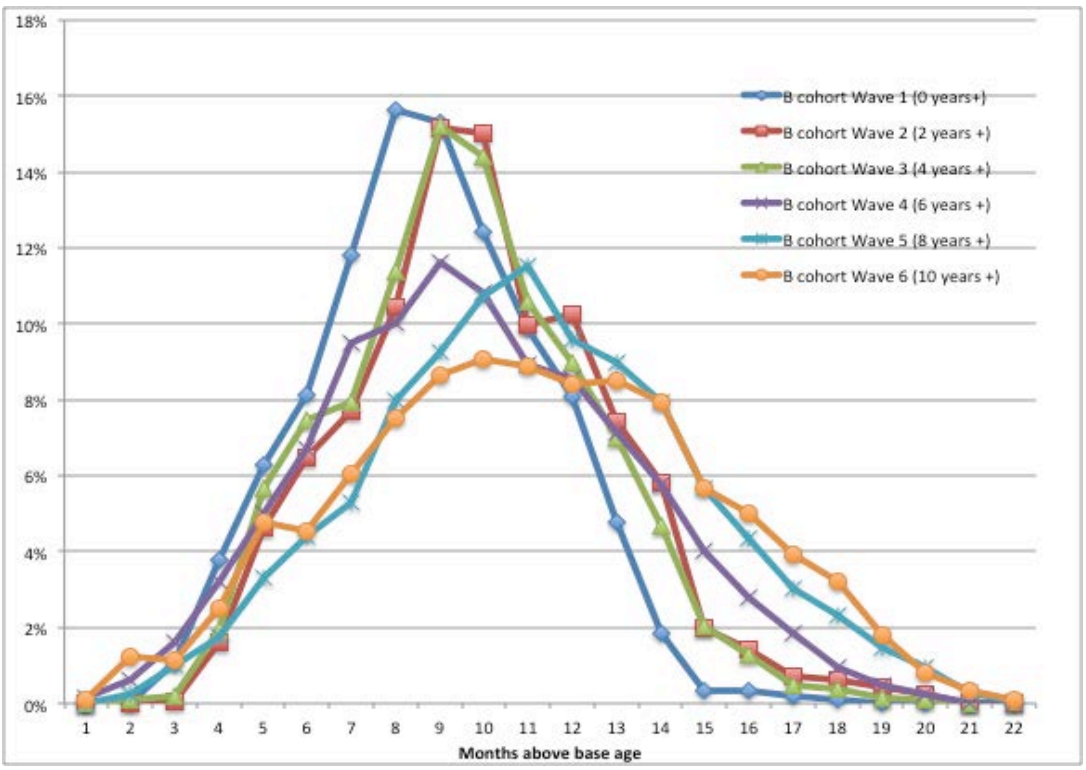


Figure 9: Age distribution of B-cohort sample at each wave

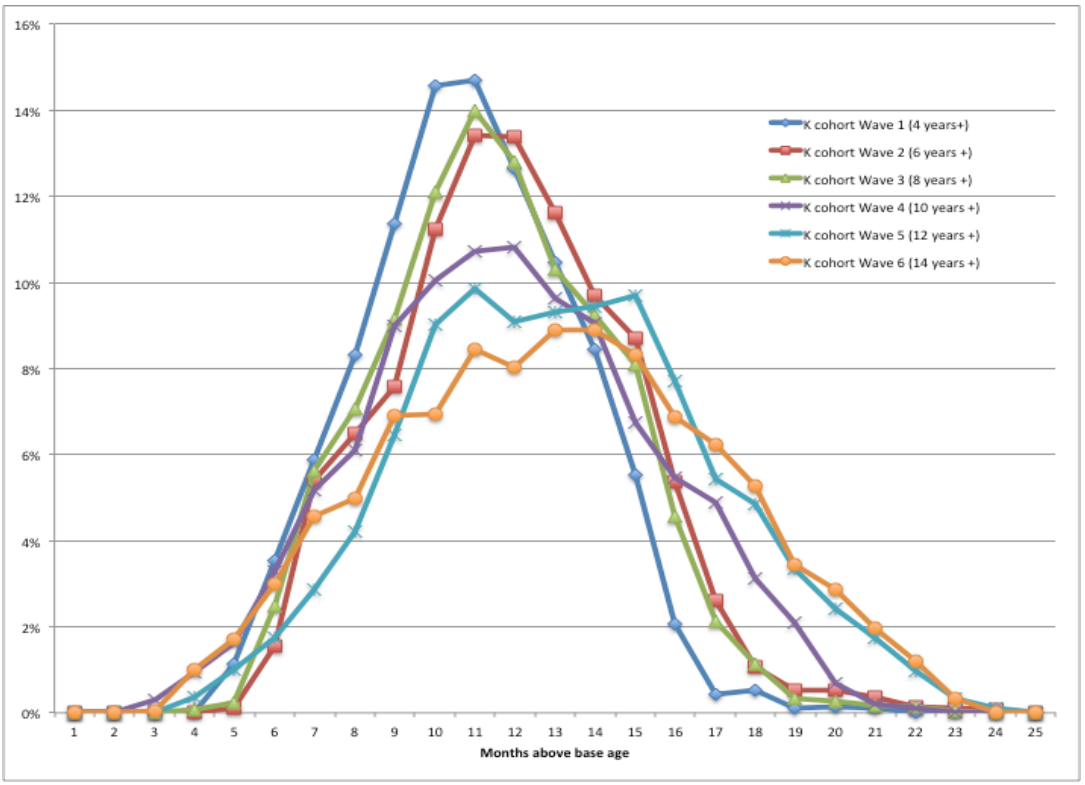


Figure 10: Age distribution of K-cohort sample at each wave

12.4 Time between interviews

Effort is made to ensure that the time between interviews is close to two years; however, in some cases this is not possible. Figures 11 and 12 show the distribution of the intervals between waves.

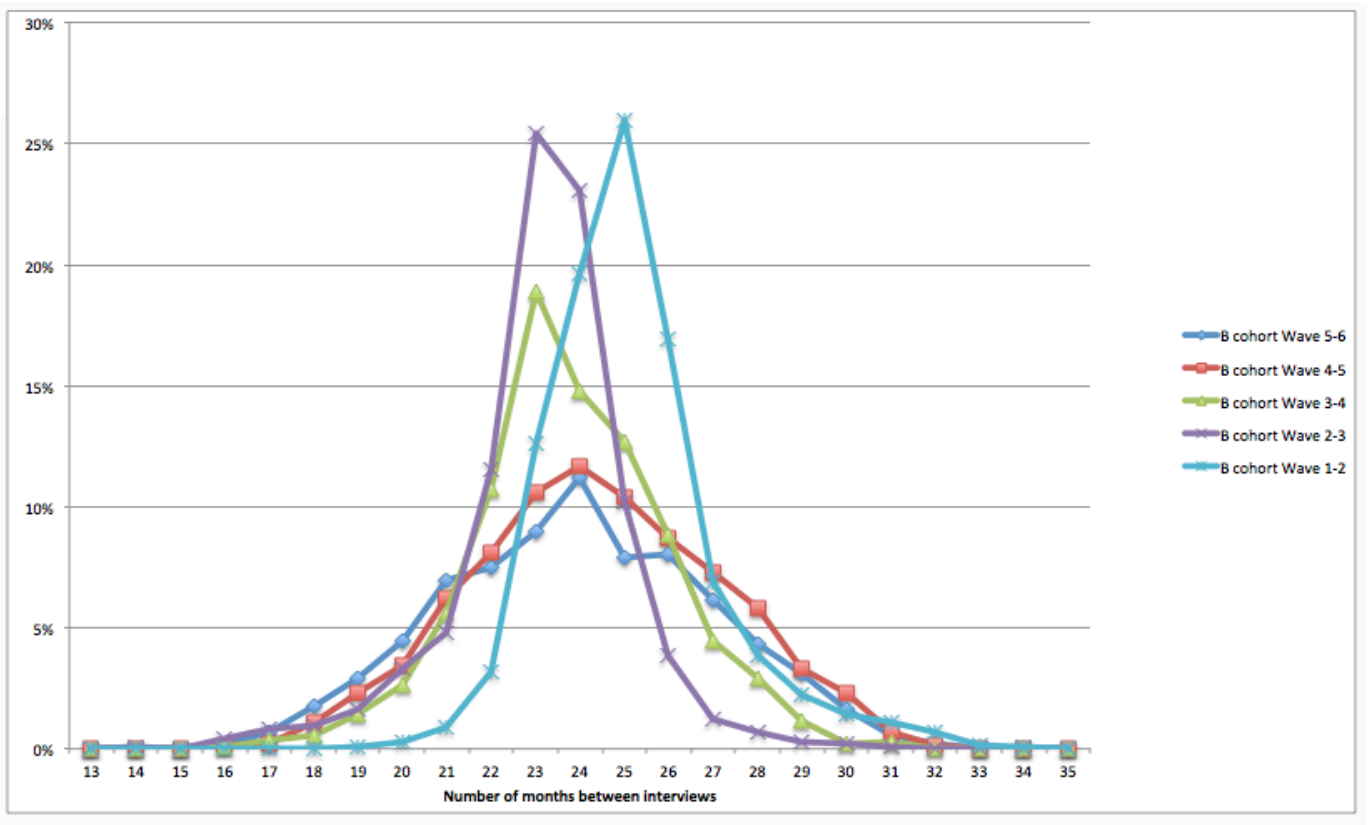


Figure 11: Distribution of time between interviews, B cohort, Waves 1 to 6

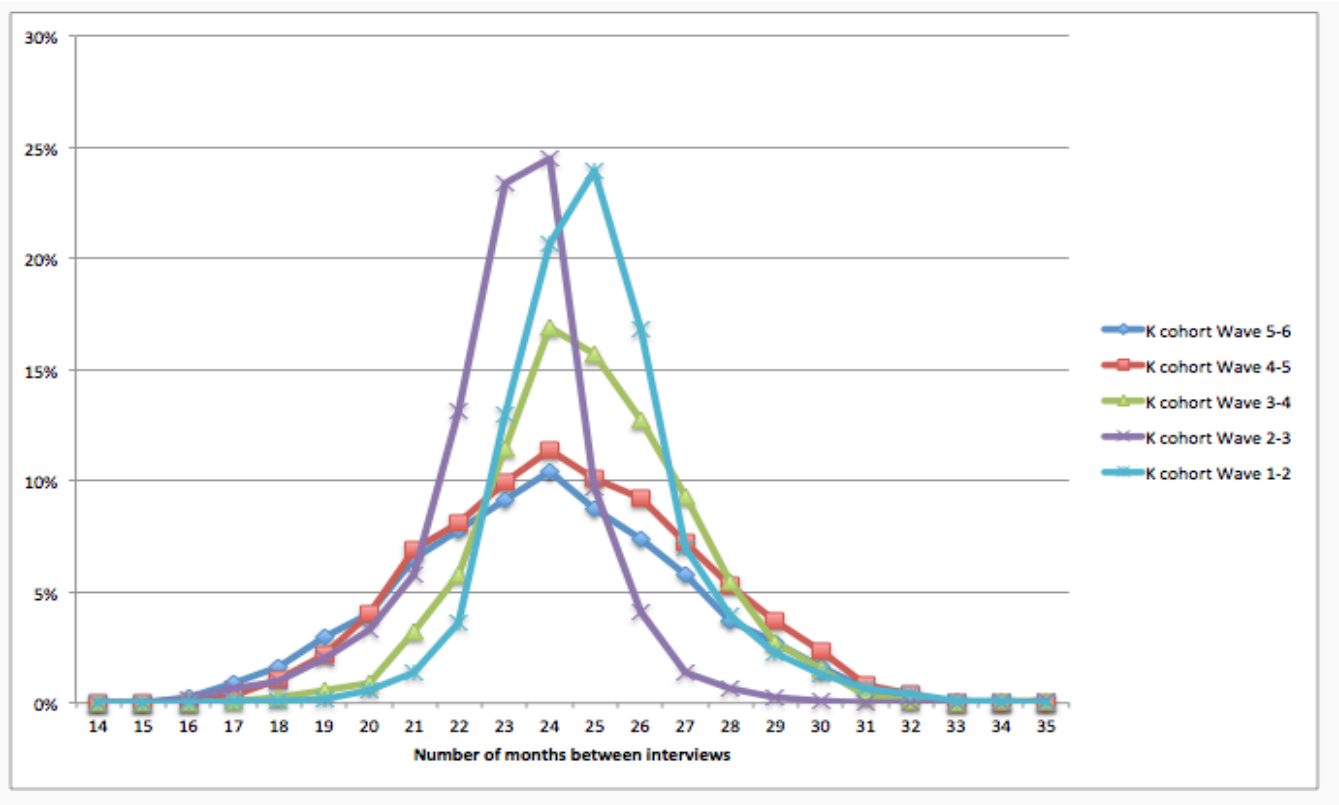


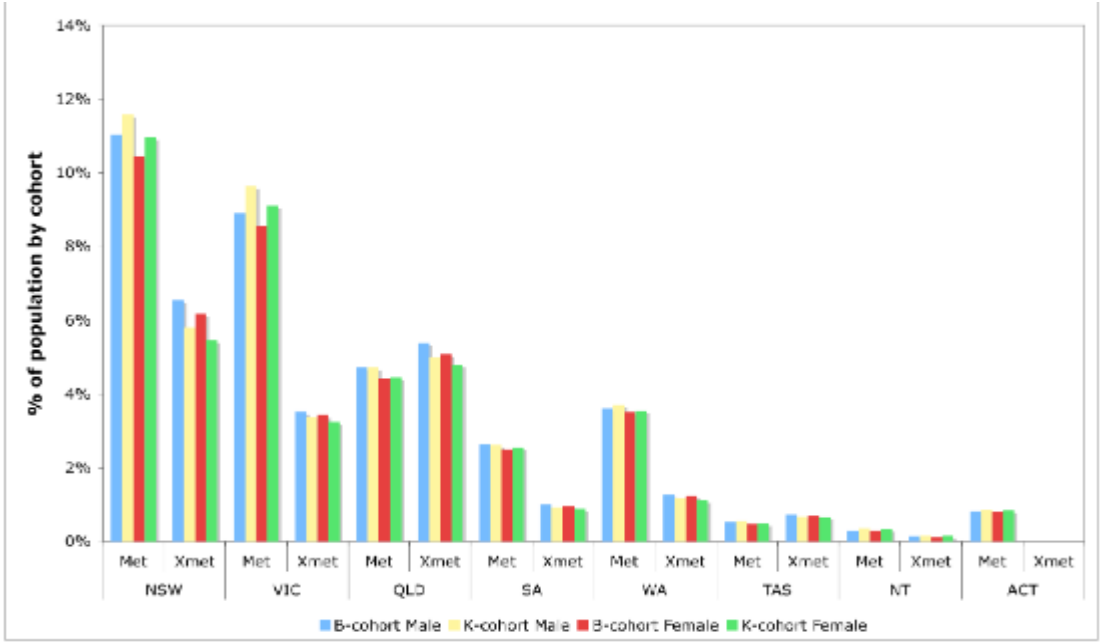
Figure 12: Distribution of time between interviews, K cohort, Waves 1 to 6

12.5 Cross-cohort comparisons

It should be noted that the two cohorts of LSAC were selected and weighted to represent similar but different populations. For the B cohort the reference population is "0 year old children in Australia in 2004 excluding those from certain remote postcodes", while for the K cohort the reference population is "4-year-old children in Australia in 2004 excluding those from certain remote postcodes". One implication of this is that the K cohort will have a greater number of children born overseas as there was more time for families to immigrate to Australia between the birth of their child and selection into the study. The 2001 census contained 4.4% of 4 year olds born overseas compared with 0.8% of 0 year olds. In comparison, the weighted percentages for these figures in LSAC at Wave 1 were 4.2% vs 0.4%.

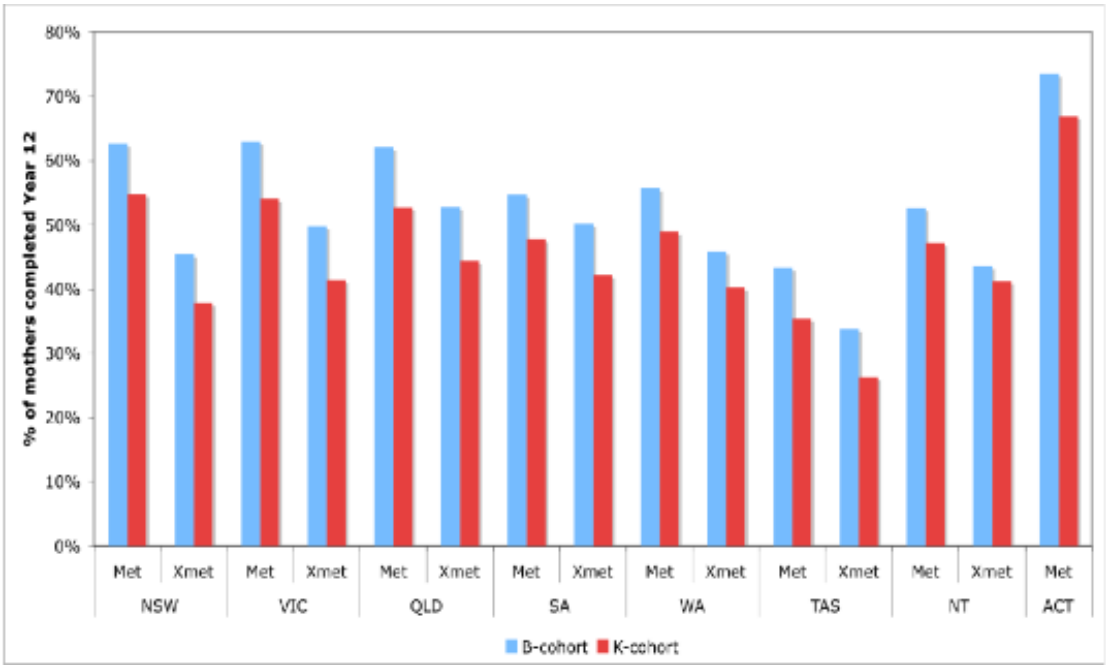
However, there are also other demographic differences between the populations that are reflected in the benchmarks used to weight the two cohorts. Figure 13 shows the population percentages in each state by part of state by gender stratum for the B and K cohorts. The B and K cohort figures generally match closely; however, the population from which the K cohort was selected was a little more likely to live in capital cities (66.5% vs 63.6%). Figure 14 shows the population proportions for mothers having completed Year 12 by state and part of state for each cohort. The B-cohort population was more likely to have completed Year 12 in every part of the country, with the ABS census figures nationally being 56.6% for the B cohort against 48.3% for the K cohort. Figure 15 shows the population proportions for mothers speaking a language other than English at home by state and part of state for each cohort. These proportions were more closely matched between the B and K cohorts.

The implications of this are that even though the two cohorts have been weighted using similar variables, it does not mean that the variables that they have been weighted on are not responsible for the differences observed between the two. For example, because the two cohorts have had non-response due to maternal education adjusted for, it does not mean they will have equal proportions of mothers who had completed Year 12 when the weights are applied. Therefore, different levels of maternal education could explain differences observed between the two samples in the educational attainment of children.



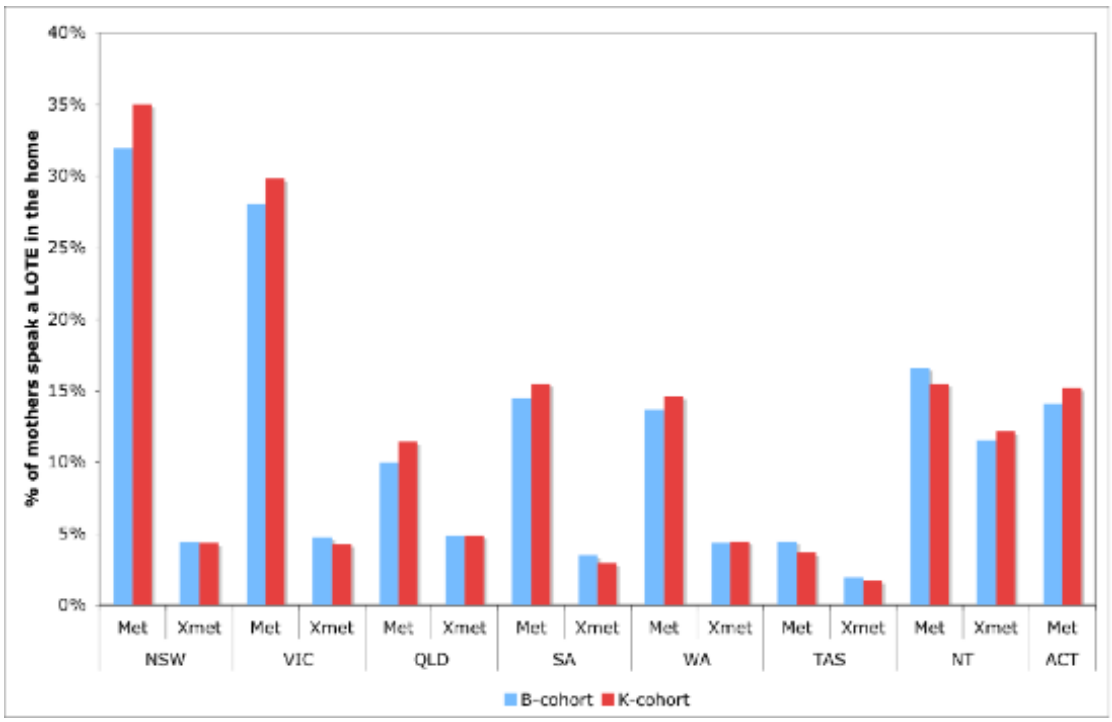
Note: There are no respondents from non-metropolitan ACT.

Figure 13: Cohort benchmarks by state, part of state and gender



Note: There are no respondents from non-metropolitan ACT.

Figure 14: Proportion of mothers who completed Year 12, cohort benchmarks by state and part of state



Note: There are no respondents from non-metropolitan ACT.

Figure 15: Proportion of mothers who speak a language other than English at home, cohort benchmarks by state and part of state

12.6 Sample characteristics

To assist in the assessment of the representativeness of the Wave 1 sample, selected characteristics were compared with ABS estimates: gender, state and region were compared with the ABS September 2004 Estimated Resident Population figures; the other characteristics were compared with (previously unpublished) population data from the ABS 2001 Census of Population and Housing (see Table 13).

Table 13: Wave 1 sample characteristics compared with ABS data

	B cohort		K cohort	
	LSAC %	ABS %	LSAC %	ABS %
Gender*				
Male	51.2	51.3	50.9	51.3
Female	48.8	48.7	49.1	48.7
Family type				
Two resident parents/guardians	90.7	88.1	86.0	82.0
One resident parent/guardian	9.3	11.9	14.0	18.0
Siblings				
Only child	39.5	36.2	11.5	12.1
One sibling	36.8	35.6	48.4	45.9
Two or more siblings	23.7	28.2	40.1	42.0
Ethnicity				
Study child Indigenous	4.5	4.3	3.8	4.3
Mother speaks a language other than English at home	14.5	16.8	15.7	17.6
Educational status				
Mother completed Year 12	66.9	56.6	58.6	48.3
Father completed Year 12	58.5	50.2	52.7	45.3
State*				
New South Wales	31.6	34.1	31.6	33.7
Victoria	24.5	24.6	25.0	23.8
Queensland	20.6	19.3	19.8	19.7
South Australia	6.8	6.8	6.8	7.2
Western Australia	10.4	9.9	10.2	10.1
Tasmania	2.2	2.3	2.7	2.5
Northern Territory	1.7	1.4	1.7	1.6
Australian Capital Territory	2.1	1.7	2.3	1.3
Region				
Capital city statistical division	62.5	63.7	62.1	62.1
Balance of state	37.5	26.3	37.9	37.9
Total	5,047		4,983	

Note: ABS data comes from the 2001 Census for families for 0 and 4 year olds, except where indicated with a *, where it is based on the September 2004 Estimated Resident Population for families of 0 and 4 year olds.

For most characteristics, the Wave 1 sample is only marginally different to the ABS data. The largest difference is in the educational status of the parents. Children with mothers who have completed Year 12 are over-represented in the sample, with proportions 10% higher than in the 2001 Census.

Other differences in the Wave 1 sample include:

- Children in lone-parent families are under-represented.

- Children with two or more siblings are under-represented and only children are over-represented in the B cohort.
- Children from an ATSI background are under-represented for the K cohort, and marginally over-represented for the B cohort.
- Children with mothers who speak a language other than English at home are under-represented.
- Children in New South Wales are under-represented.

Table 14 shows the number of children in the Wave 1 sample with selected characteristics, and gives the Waves 2 to 6 response rates for children with these characteristics. As can be seen in the table, the greatest sample loss has been from Indigenous families and families where Parent 1 speaks a language other than English at home.

Table 14: Response rates at Waves 2 to 6 by selected sample characteristics

	Wave 1 *N*		% responding at Wave 2		% responding at Wave 3	
Characteristics	B	K	B	K	B	K
Full sample	5,107	4,983	90.2	89.6	85.9	86.9
Study child male	2,610	2,537	90.0	89.8	86.2	87.2
Study child female	2,497	2,446	90.3	89.4	85.5	86.6
Study child Indigenous	230	187	78.3	81.8	64.8	66.3
Mother speaks language other than English	740	778	83.9	83.8	75.0	76.6
Mother did not complete Year 12	1,688	2,044	84.8	86.5	78.8	81.7
Father did not complete Year 12	1,890	2,016	90.0	90.1	85.9	87.0
New South Wales	1,615	1,573	90.3	90.1	84.4	86.3
Victoria	1,251	1,245	88.4	86.3	85.1	86.0
Queensland	1,054	988	91.4	90.8	88.0	87.2
South Australia	347	339	91.1	89.4	88.2	86.7
Western Australia	533	507	89.7	91.5	83.9	87.6
Tasmania	113	136	90.3	94.1	92.0	91.2
Northern Territory	87	82	90.8	89.0	83.9	87.8
Australian Capital Territory	107	113	97.2	94.7	95.3	94.7
Capital city statistical division	3,194	3,095	90.6	89.3	86.2	86.8
Balance of state	1,913	1,888	89.5	90.0	85.4	87.2

	% responding at Wave 4		% responding at Wave 5		% responding at Wave 6	
Characteristics	B	K	B	K	B	K
Full sample	83.1	83.7	80.0	79.4	73.7	71.0
Study child male	83.9	84.1	80.3	79.7	73.9	70.9
Study child female	82.2	83.2	79.6	79.1	73.4	71.0
Study child Indigenous	63.0	63.1	60.4	60.4	46.1	44.4
Mother speaks language other than English	72.0	71.1	68.6	66.1	61.1	58.5
Mother did not complete Year 12	74.4	78.1	70.1	72.6	61.4	62.2
Father did not complete Year 12	83.6	84.9	79.7	80.9	73.0	71.6
New South Wales	81.8	81.8	79.8	78.2	71.2	70.2
Victoria	81.9	83.1	76.6	76.7	71.5	68.1
Queensland	84.3	84.0	82.4	80.9	75.4	71.9
South Australia	85.9	83.2	81.0	79.6	76.1	70.5
Western Australia	81.6	86.0	78.6	81.1	75.0	73.0
Tasmania	92.9	90.4	91.2	87.5	88.5	83.1
Northern Territory	80.5	89.0	81.6	86.6	79.3	72.0
Australian Capital Territory	93.5	92.0	89.7	89.4	85.0	82.3

Capital city statistical division	82.9	82.8	79.9	78.7	74.1	70.4
Balance of state	83.3	85.0	80.1	80.5	73.0	72.0

13 User support and training

User training sessions are offered by AIFS to further develop the information provided in the user manual and to allow users to interact with the LSAC Data Management team and benefit from their knowledge and experience with the data. These sessions consist of an introduction to LSAC and the newly released datasets including:

- study methodology
- introduction to the datasets
- issues for data analysts (e.g., weighting, clustering, confidentialisation)
- variable naming
- user resources (e.g., data dictionary, labelled questionnaires).

See the LSAC website for details on when training sessions are being offered.

13.1 Online assistance

An email alert list is used to convey key information and updates to users. Important information distributed via the email alert list is also stored in the data access area of the *Growing Up in Australia* website. This area contains:

- all reference material made available to users (in downloadable form)
- Excel data dictionary
- critical updates and alerts as distributed through the email alert list
- updates on data-user workshops.

13.2 Getting more information

More information on *Growing Up in Australia* and its progress can be found on the LSAC website. <www.growingupinaustralia.gov.au/index.html>

Further enquiries can be directed to aifs-lsac@aifs.gov.au or by contacting the:

LSAC Data Manager
Australian Institute of Family Studies
Level 20/485 La Trobe Street
Melbourne VIC 3000
Tel: +61 3 9214 7803
Fax: + 61 3 9214 7839

References

Chambers, R. L., & Skinner, C. J. (Eds.). (2003). *Analysis of survey data*. Chichester: Wiley.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*, 3rd Edition. New York: Wiley.

National Childcare Accreditation Council. (2003). *OSHCQA quality practices guide* (1st ed.). Sydney: NCAC.

National Childcare Accreditation Council. (2003). *QIAS validation report* (2nd ed.). Sydney: NCAC.

National Childcare Accreditation Council. (2004). *FDCQA quality practices guide* (2nd ed.). Sydney: NCAC.

National Childcare Accreditation Council. (2006). *QIAS quality practices guide* (1st ed.). Sydney: NCAC.

Rice M.L., Hoffman L., & Wexler K. (2009). *Judgments of omitted BE and DO in questions as extended finiteness clinical markers of specific language impairment (SLI) to 15 years: A study of growth and asymptote.* Journal of Speech, Language, and Hearing Research. 52:1417–1433.

Rowe, K. (2006). *The measurement of composite variables from multiple indicators: Applications in quality assurance and accreditation systems—childcare*. Background paper prepared for the National Childcare Accreditation Centre, Sydney.

Bibliography

The following publications provide more information on techniques for analysis longitudinal and survey data:

Australian Bureau of Statistics. (1996). *Women's safety Australia: User guide*. Canberra: ABS.

Baxter, J. (2007). *Children's time use in the Longitudinal Study of Australian Children: Data quality and analytical issues in the 4-year-old cohort* (LSAC Technical Paper No. 4). Melbourne: Australian Institute of Family Studies.

Chambers, R. L., & Skinner, C. J., (Eds.). (2003). *Analysis of survey data*. Chichester: Wiley.

Cusack, B., & Defina, R. (2013). *Wave 5 weighting and non response* (LSAC Technical Paper No. 10). Canberra: Australian Bureau of Statistics.

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

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376-382.

Deville, J. C., Särndal C. E., & Sautory, O. (1993). Generalised raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*, 1013-1020.

Kalton, G. (1983). *Compensating for missing survey data* (Research report series). Ann Arbor, MI: Institute for Social Research, University of Michigan.

Lepkowski, J. M. (1989). Treatment of wave nonresponse in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton & M. P. Singh (Eds.), *Panel Surveys* (pp. 348-374). New York: Wiley.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications* (3rd ed.). New York: Wiley.

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

National Childcare Accreditation Council. (2005). *FDCQA quality practices guide* (3rd ed.) Sydney: NCAC.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society Series B*, *60*, 23-40.

Sanson, A., Misson, S., & The LSAC Outcome Index Working Group. (2006). *Summarising children's wellbeing: The LSAC Outcome Index* (LSAC Technical Paper No 2). Melbourne: Australian Institute of Family Studies.

Sanson, A., Nicholson, J., Ungerer, J., Zubrick, S., Wilson, K., Ainley, J. et al. (2002). *Introducing the Longitudinal Study of Australian Children* (LSAC Discussion Paper No. 1). , Melbourne: Australian Institute of Family Studies.

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

Skinner, C. J., & Holmes, D. J. (2003). Random effects models for longitudinal survey data. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of Survey Data* (pp. 205-218). Chichester: Wiley.

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

Soloff, C., Lawrence, D., Misson, S., & Johnstone, R. (2005). *Wave 1 weighting and non-response* (LSAC Technical Paper No. 3). Melbourne: Australian Institute of Family Studies.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using Multivariate Statistics* (2nd ed.). New York: Harper and Row.

Wolter, K. (1984). *Introduction to Variance Estimation*. New York: Springer.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Information on related studies:

Fergusson, D. M., Horwood, L. J., Shannon, F. T., & Lawton, J. M. (1989). The Christchurch Child Development Study: A review of epidemiological findings. *Pediatric and Perinatal Epidemiology*, *3*, 278-303.

Freidin, S., Watson, N., & Wooden, M. (2002). *HILDA survey coding framework: Confidentialised data* (HILDA project technical paper series). Melbourne: The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Frick, J. R., & Haisken-DeNew, J. P. (2001). *Structuring the HILDA panel: Considerations and suggestions* (HILDA project discussion paper series). Melbourne: The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Henstridge, J. (2001). *The Household Income and Labour Dynamics in Australia (HILDA) survey: Weighting and imputation* (HILDA project discussion paper series). Melbourne: The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

National Longitudinal Survey of Children and Youth (NLSCY). (1999). *Overview of survey instruments for 1998-99 data collection cycle 3*. Catalogue no. 89FOO78XPE, No. 3. Ottawa: Statistics Canada.

Watson, N., & Fry, T. R. L. (2002). *The Household Income and Labour Dynamics in Australia (HILDA) survey: Wave 1 weighting* (HILDA project technical paper series). Melbourne: The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Watson, N., & Wooden, M. (2002). *The Household Income and Labour Dynamics in Australia (HILDA) survey: Wave 1 Survey Methodology* (HILDA project technical paper series). Melbourne: The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Willms, D. (Ed.) (2002). *Vulnerable children*. Edmonton: University of Alberta Press.

Wooden, M. (2001). *Design and management of a household panel survey: Lessons from the international experience* (HILDA project discussion paper series).Melbourne: The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Wooden, M., & Watson, N. (2000). *The Household Income and Labour Dynamics in Australia (HILDA) survey: An introduction to the proposed survey design and plan* (HILDA project technical paper series). Melbourne: The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.