The Longitudinal Study of Australian Children:
An Australian Government initiative

Data Issues
Waves 1 to 6

2015

Contents

List of tables

List of figures

1 Introduction

This paper provides a summary of data-related issues that have emerged over the life of *Growing up in Australia*: The Longitudinal Study of Australian Children (LSAC). The chapters were initially published on the LSAC website as a series of Issues Papers designed to assist users of the LSAC data as they undertake research and analysis of the LSAC datasets.

The paper is to be used in conjunction with the Data User Guide available on the LSAC website.

2 Cleaning of time use diary data

2.1 Background

The LSAC time use diary (TUD) is a diary consisting of 96 15-minute time intervals or bubbles with pre-coded activity (e.g. sleeping, eating, bathing) and context (e.g. where they were and who they were with) information. Parents are asked to mark which of the pre-coded activities were done during each of the 96 time intervals. The diary begins at 4 am. Thus, time interval 1 is from 4 am to 4:14 am, time interval 2 is from 4:15 am to 4:29 am, etc. For the B cohort at Wave 1 there were 22 pre-coded activities, five context locations and seven "who else was present" context options. Additionally, the diarists were asked whether they had paid for the activity that the child was doing. For the Wave 1 B cohort the total matrix size was 3,360, consisting of the 35 (activities and context descriptors) by 96 time intervals. The Wave 1 K cohort had 26 pre-coded activities. Otherwise, the diary was the same as for the B cohort. The matrix size for the K cohort was 3,744.

The data entry used scanning technology. For Wave 1, few checks were made at the time of data entry and subsequently it has been found that the scanner was sensitive to rub outs and other marks that appeared in the bubbles on the paper files. This resulted in false data (false positives) that exists in the electronic data files but does not exist on the paper files.

For Wave 2 various procedures were implemented to ensure that these problems did not recur. These procedures involved changes to the data capture and data validation stages.

To reduce problems associated with capturing the data, changes were made to the scanner settings. Through the Intelligent Forms Processing (IFP) system it is possible to define the minimum character/mark size that will be registered by the system. Testing of TUD capture confirmed that oversensitivity of scanning equipment can produce a high rate of false positive responses on the TUD. Following iterative testing of LSAC dress rehearsal (DR) TUD forms, it was determined that the character size for the TUD "bubbles" should be increased from 2 x 2 pixels to 2 x 5 pixels. Testing showed that this setting allowed the IFP system to disregard very small specs of dust, etc. (thereby greatly reducing false positives) without any impact on the false negative rate.

A second setting that impacts on the registration of marks is the size of field which is "scanned" for a character/mark. Often this field size is expanded slightly beyond the expected capture area to allow marks falling slightly outside the response box to be registered. However, in the case of the TUD, the extremely close arrangement of the response bubbles meant that such an expansion led to false positives from slight (and unintended) continuation of marks beyond one response bubble but not quite into the subsequent one. For this reason, horizontal margins for the capture area have been strictly limited to the intended response area (i.e. the border of the response bubble). However, vertical (top/bottom) margins of 6 pixels outside the response bubble have been retained to ensure that marks made slightly above/below a response bubble are captured.

Following capture, forms are forwarded for inspection and repair by a trained operator. The process outlined below is performed on all TUD forms, with the majority expected to contain at least one response mark which will need to be investigated.

The first repair process conducted on scanned forms is the on-screen inspection of mosaics of scanned response marks known as carpets. Carpets display images of all marks from the same form that have been recognised by the system. Depending on the system's confidence in the validity of a particular response, the mark will be displayed in a green, yellow or red shaded box. At this stage the operator is able to confirm or correct a response or, alternatively, select responses for further investigation through the form process (see below).

In the example below (Figure 1) the carpet displays images of eight response marks, all of which were confidently identified by the system as valid responses. In two cases the respondent has clearly attempted to correct a response by crossing out the original mark. While the Optical Mark Reader (OMR) scanner is unable to distinguish these responses from the valid responses, the analysis of carpets gives the operator an opportunity to correct the data, thereby avoiding potential false positives. Operators conducting repair of LSAC forms are trained to examine carpets for these types of responses.

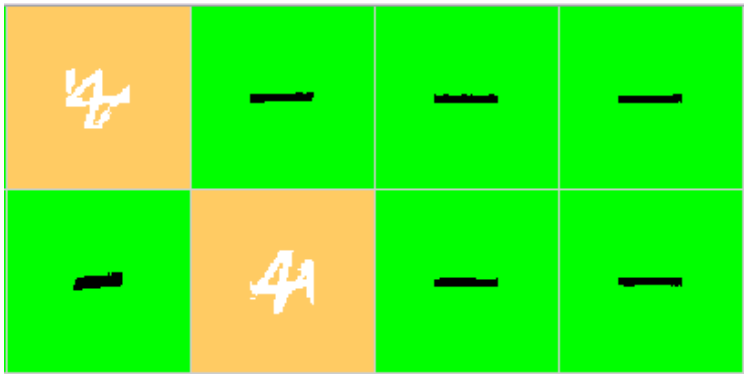

Figure 1: Example of "carpets" from time use diary scanning

Forms containing at least one mark queried by the system or the operator progress to the forms stage of repair. At this stage the operator is able to see the queried response in the context of the form and other responses.

In the test example below (Figure 2) the operator has queried the two marks highlighted in red during the earlier examination of carpets for this form. The ability to further examine these marks then enables the operator to make a more informed decision regarding its validity.

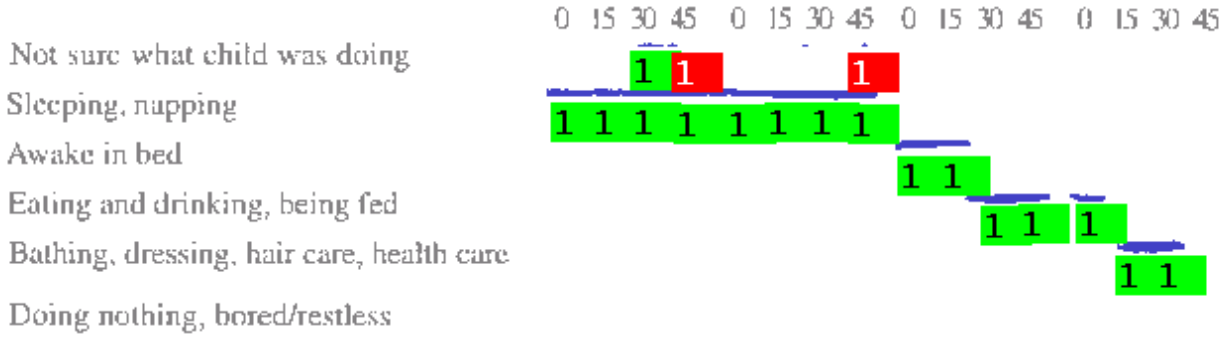

Figure 2: Example of in-context diary data display

The processes outlined above have been developed as a result of the experience of the Australian Bureau of Statistics (ABS) in the capture of LSAC TUD forms during the Wave 2 dress rehearsal and more recent testing of final forms. It should be noted, however, that these processes do not address data quality issues associated with respondent error nor will they fully overcome capture difficulties associated with formatting features of the TUD such as the extremely close arrangement of response bubbles and the sheer volume of information recorded on each page.

The rest of this section reports on the extent of the Wave 1 false positives and provides a description of attempts to electronically remove the false positives, as well as other measures to improve data quality. A number of strategies were used to recode the data, working off the premise that most implausible data combinations are likely to be false positives. For example, if it is late in the evening or early in the morning, a child is likely to be either in bed or asleep, not simultaneously sleeping and walking (not for extended

periods of time at least). The rules for the electronic recodes are outlined in more detail below. These corrections were applied only to the early wave TUD data.[1]

In recoding the data it is important that the amount of real data being incorrectly removed (false negatives) is minimised. It is expected that this incorrect coding was most likely to occur when there were transitions between activities. In order to protect against this, sequences of events were often considered. That is, comparisons were often made with the preceding and following time interval. However, it must be realised that from time to time diarists unintentionally provide information on implausible events.

In addition to the recoding of false positives, other data cleaning strategies or imputations are employed. These recodes are potentially important, as it is common in time use analysis to exclude data with more than 90 minutes of missing data. Thus, if the number of missing bubbles can be minimised, less data is lost. These imputations are performed on the data from all 3 waves.

2.2 Assessing the extent of the false positives

2.2.1 K cohort

In order to estimate the extent of the false positives and to ascertain whether corrections could be made electronically, a random sample was drawn from both the B cohort (n = 51) and K cohort (n = 49). One diary was excluded from the B cohort as the diary was returned blank and two diaries were excluded from the K cohort, one because only one activity was given for the whole diary and the other because it did not match the electronic file at all. These forms were manually checked against the electronic records so that false positives were identified and given a unique code on the electronic record. These files are known throughout this report as the "corrected files". The files of these cases before they are checked will be known as the "original files".

A summary of false positives by data type for the K cohort is presented in Table 1. Over the random sample of 47 children and over the 96 time intervals, there were 16,248 positive responses in the original file. This file had 882 "extra" bubbles or units of data than was provided in the "corrected file". Thus, the false positive rate is 6% (876/(876 + 15366)*100). It should be noted that there was one diary that had a false positive rate of 30%, while the next highest figure was 14%. If these cases were to be excluded, the false positive rate would drop to 5%.

For the K cohort, the highest aggregate false positive rate was for the "with whom" context data (7%), while for the activity data and "where" context data, the false positive rate was around 5%. The light diary also asked whether someone was paid for the activity. There was only one false positive associated with this data.

The general trend was that the more real data there was the greater the number of false positives. This is not surprising given that much of the false positive data was due to rub-outs.

[1] Note that at the time of the release of Wave 2 data the Wave 2 TUDs also had these processes applied to them in order to maximise consistency. For the release of Wave 3 data this decision was reversed. It was felt that the corrections excluded some combinations of activities that were unlikely but possible, particularly as the children became older (e.g. sleeping outside).

Table 1: Summary of false positives based on comparing original and corrected file of 47 cases (K cohort, Wave 1).

Source	True positives	False positives	False positive rate (%)
Total	15,366	876	5.7
Activity	5,241	274	5.2
Where	3,738	195	5.2
Who	6,231	406	6.5
Paid	156	1	0.6

2.2.2 B cohort, Wave 1

A summary of false positives by data type for the B cohort is given in Table 2. Over the random sample of 50 children over the 96 time intervals, there were 17,703 units or bubbles of data in the original file. This file had 723 false positive values or "extra" bubbles or units of data than was provided in the "corrected file". Thus, the total false positive rate was 4% (723/17703*100).

Table 2: Summary of false positives based on comparing original and corrected file of 50 cases (B cohort, Wave 1)

Source	True positives	False positives	False positive rate (%)
Total	16,980	723	4.1
Activity	5,898	311	5.0
Where	4,235	107	2.5
Who	6,732	302	4.3
Paid	115	3	2.5

2.3 Recoding to reduce false positives

2.3.1 K cohort

A number of recodes were experimented with to reduce the rate of false positives. Table 3 gives a summary of the impact of electronic recodes on the original file in terms of both reducing the number of false positives as well as introducing false negatives into the data. These recodes reduced the false positive for the test file to 5%. The recodes are described in greater detail below.

Table 3: Summary of electronic corrections made to original file of 47 cases (4-5 year olds)

	Imputed true negatives[a]	Imputed false negatives[b]
If sleeping (recode other activities = 0)		
Early morning (4 am to 9 am)	64	6
Late evening (9 pm to 4 am)	20	6
If awake in bed (recode other activities = 0)	1	0
If one activity and implausible context data		
Travel (recode home inside or alone)	12	11
Walk/ride (recode inside a home)	11	1
Television and computer (recode outside)	1	0
If sleeping or awake in bed (recode outside)	12	0
If at day care centre outside plausible hours and give other location (recode day care)	4	0
Total activities and context data	125	24

Notes: [a] Imputed true negatives are those cells which were imputed as "0" where the corrected file had indicated that the positive response was a false positive. [b] Imputed false negatives are those cells which were imputed to "0" where the corrected file indicated that the positive response was a true positive.

Correction 1: Being asleep and doing other activities at the same time

In order to reduce the likelihood of recoding a transitional phase, the child had to be asleep in a given interval and in the preceding and following interval. Three separate time periods were tested: morning (4 am to 9 am), nighttime (9 pm to 4 am) and daytime (9 am to 9 pm).

For the morning and nighttime periods, if an activity occurred simultaneously with sleep, non-sleep activity time was coded as zero. The recoding of activities occurring simultaneously with sleep in the morning and in the evening was relatively successful, with an aggregate number of 84 correct recodes and 12 incorrect recodes. In the full Wave 1 file (i.e. *N* = 7,449) this resulted in 23,216 recodes of positive responses. This correction was not performed on the Wave 2 data (*N* = 6,906); however, if it had been it would have only resulted in 4,498 recodes. This provides further evidence that this correction removed many more false positives than true ones.

In the period between 9 am and 9 pm, children were most likely to be periodically transitioning between sleep and other activities. Attempting to recode activities occurring simultaneously with sleep, in this period, yielded no corrections to false positives and resulted in eight true positives being recoded. Alternatively, an attempt to recode sleep resulted in recoding four false positives and five true positives. Given that both these alternatives resulted in more incorrect recodes than correct ones, neither was performed on the main data file.

Correction 2: Being awake in bed and doing other unlikely activities at the same time

Other unlikely activities are defined as:

- bathing, dressing, hair care, health care
- using computer/computer games
- walking for travel or fun
- riding bicycle, trike, etc. (travel or fun)
- other exercise—swim/dance/run about
- travelling in pusher or on bicycle seat
- travelling in car/other household vehicle
- travelling on public transport, ferry, plane
- taken places with adult (e.g. shopping)
- organised lessons activities.

If the children were doing these activities as well as being awake in bed, other activities were coded as zero. Again, in order to reduce the likelihood of recoding a transitional phase, the child had to be awake in bed in a given interval, while in the following and preceding interval they had to be either asleep or awake in bed.

The impact of recoding activities occurring simultaneously with the child being awake in bed was relatively minor, with only one false positive being recoded in the 47 diaries in the original file, and 1,225 positive responses being recoded in the full file. This is not surprising given that 4-5 year old children are not often awake in bed for long periods of time unless they are ill or are having trouble getting to sleep. In Wave 2, 1,077 positive responses would have been recoded due to this correction.

Correction 3: A child cannot be travelling and be inside at home or be alone

A child cannot be travelling (travelling in a pusher/ travelling in a car/ travelling on public transport/taken places with an adult) and be simultaneously at home inside (or in someone else's home) or be alone, if travelling was their sole activity for the time period. Recoding the context data as 0 where this occurred resulted in slightly more false positives being altered than true ones. In the full file this resulted in 4,111 positive responses being recoded. In Wave 2, 1,359 responses would have been recoded.

Correction 4: A child cannot be walking/riding and be inside a home

A child cannot be walking for travel or fun or riding a bike or trike, etc. and be inside their own or someone else's home if this is their only activity for the time period. Recoding all incidences of being inside as zero resulted in many more false positives being altered than true ones. In the full file this resulted in 1,996 positive responses being recoded. In Wave 2, only 282 responses would have been recoded.

Correction 5: A child cannot be watching television or using a computer and be outside

A child cannot be watching television or using the computer and be outside. Recoding being outside as zero in this situation resulted in only one false positive correction for these 47 diaries but no alterations to true positives. In the full file this correction resulted in the recoding of 430 positive responses. In Wave 2, 157 responses would have been recoded.

Correction 6: A child cannot be sleeping or awake in bed and be outside

If a child was awake in bed or asleep and this was their only activity they cannot be outside. While this does exclude any children who were camping (assuming a tent doesn't count as indoors), in the test cases available this correction eliminated 12 false positives without altering a true positive. In the full file this resulted in 2,459 positive responses being recoded. In Wave 2, 535 responses would have been recoded.

Correction 7: A child cannot be at a day care centre/play group outside the hours of 7 am to 7 pm

If a response was given outside of these hours it was recoded to zero. In the original file this resulted in four corrections to false positives without creating any false negatives, while in the full file 1,417 responses were recoded by this correction. In Wave 2, 610 responses would have been recoded.

2.3.2 B cohort

Table 4 gives a summary of the impact of electronic recodes on the original file for the infants in terms of both reducing the number of false positives as well as introducing false negatives into the data. In summary, the recoding resulted in a reduction of 121 of the 723 false positive data, with little creation of false negatives. As a result of these recodes, the false positive rate fell from 4% to 3%. The recodes are described in greater detail below.

Table 4: Summary of electronic corrections made to original file of 50 cases (B cohort, Wave 1)

	N	Imputed true negatives[a]	Imputed false negatives[b]
If sleeping (recode other activities = 0)			
4 am to 7 am	50	28	8
If alone/sleeping then can't be with others			
4 am to 3 pm	50	38	15
10 pm to 4 am	50	22	0
Travelling and at home or alone	50	27	9
If one activity is breastfeeding, bathing, being held or read to, recode alone	50	4	1
If at day care centre outside of the hours of 7 am to 7 pm and give other location (recode day care)	50	2	0
Total activities and context data	50	121	33

Notes: [a] Imputed true negatives are those cells that were imputed as "0" where the corrected file had indicated that the positive response was a false positive. [b] Imputed false negatives are those cells that were imputed to "0" where the corrected file indicated that the positive response was a true positive.

Correction 1: Being asleep and doing other activities at the same time

Children should not be asleep and also be active in another activity in the same time interval unless the child was in transition between activities. For this recode, activities that were indicated in the same time period of sleep were recoded for intervals where the child was also asleep in the preceding and following interval. This recode was tried for a number of different time periods but was only successful at the start of the day between 4 am and 7 am, where it recoded 28 false positives and 8 true ones. In the full file (N = 7,782) this recode resulted in 11,278 positive responses being altered. As for the K cohort, these corrections were not repeated in Wave 2; however, if they had of been, only 1,464 responses would have been recoded, suggesting that most Wave 1 responses recoded were false positives.

Correction 2: Being asleep alone and with someone at the same time

If the child was sleeping alone in a time period as well as the one preceding and following it, all other data in the "in the same room as" section was recoded to zero. The only time period this didn't work for was the evening between 3 pm and 10 pm, so this period was excluded from this recode. Outside of these times it resulted in 60 corrections to false positives while introducing only 15 false negatives. In the full file this recode resulted in 29,016 responses being altered. In Wave 2, 8,618 responses would have been recoded.

Correction 3: A child cannot be travelling and be inside at home or be alone

A child cannot be travelling (travelling in a pusher/ travelling in a car/ travelling on public transport/taken places with an adult) and be simultaneously at home inside or be alone. A child was identified as travelling if they are travelling in a given interval and the preceding and following interval. In this situation the "at home" or "alone" response would be removed. This correction removed 27 false positives while introducing nine false negatives. In the full file, it led to 6,303 positive responses being removed. In Wave 2, 1,098 positive responses would have been removed.

Correction 4: Being alone and with improbable activities

If a child's only activities are breastfeeding, being held, having personal grooming tasks performed, or being read a story or talked or sung to, any response that the child was alone for the period was recoded to zero. Correcting this removed four false positives and produced one false negative, while in the full file this recode lead to the removal of 1,031 responses of "alone". In Wave 2, 420 responses would have been removed.

Correction 5: Being at a day care centre/playgroup at improbable hours

Any response indicating that the child was at a day care centre outside the hours of 7 am to 7 pm was recoded to zero. This recode corrected two false positives and produced no false negatives, while in the full file, 593 positive responses were removed. In the Wave 2 file, 301 responses would have been removed.

2.4 Coding to improve data quality

A number of further recodes were undertaken to improve other aspects of data quality, such as reducing missing or contradictory data.

2.4.1 B and K cohorts

These operations were performed on all three waves of diary data for both cohorts.

Improvement 1: Recoding not sure when other activities given in the same time interval

Ideally, respondents should only have given "not sure" as a response if they were unable to report any of the child's activities in a 15-minute block. Where this has happened the "not sure" response was coded to zero. For Wave 1, this removed 12,770 "not sure" responses from the K-cohort file and 10,026 "not sure" responses from the full infant file. For Wave 2, these figures were 4,678 for the K cohort and 3,717 for the B cohort and, in Wave 3, they were 5,380 for the K cohort and 3,759 for the B cohort.

Improvement 2: Imputing not sure or missing activity data as sleep and the early morning

If the parent was not sure of what the child was doing or activity data was missing in the early morning (4 am to 9 am) and the sequence of not sure/missing ended with either the child being awake in bed or sleeping, the not sure/missing was recoded as sleep. In Wave 1, these changes created an extra 984 sleep responses in the full K-cohort file and 1,292 extra sleep responses in the full B cohort file. In Wave 2, these figures were 1,036 for the K cohort and 1,002 for the B cohort and, in Wave 3, they were 943 for the K cohort and 566 for the B cohort.

Improvement 3: Imputing not sure or missing activity data as sleep at nighttime

If the parent was not sure of what the child was doing or activity data was missing at nighttime (9 pm to 4 am) and this sequence began following the child being either awake in bed or sleeping, the not sure/missing data was recoded as sleep. This created an extra 2,540 sleep responses in the full K-cohort file and 4,101 in the full B-cohort file. In Wave 2, the figures were 2,681 for the K cohort and 3,933 for the B cohort and, in Wave 3, they were 2,229 for the K cohort and 2,328 for the B cohort.

Improvement 4: Other missing data

If there was a single time period with missing activity data and the child remained in the same location, then either the activity before or after the missing bubble was randomly allocated to the missing bubble. This improvement imputed activities in 2,517 time periods in the full K-cohort file and 3,022 time periods in the B-cohort file. For Wave 2, these figures were 1,956 for the K cohort and 2,389 for the B cohort and, in Wave 3, they were 1,553 for the K cohort and 1,767 for the B cohort.

Improvement 5: Missing "who" information in child care

If a child's "where" information includes "day care centre/playgroup", it can reasonably be assumed they are in the presence of other children and other adults when alternative information is missing. This improvement imputed data in 17,294 time periods in the full K-cohort file and 3,169 time periods in the B-cohort file. As might be expected given the rise in time in non-parental care for the children as they get older, these numbers were higher in Wave 2 with 35,097 time periods for the K cohort and 12,712 for the B cohort and, in Wave 3, they were 23,045 for the K cohort and 15,835 for the B cohort.

2.5 Exclusion of cases

It is common practice when analysing time use diary data to exclude cases with poor quality data, often indicated by rules of thumb such as more than 90 minutes of missing information (e.g. Egerton and Gershuny, 2004; Fisher, 2002). The LSAC time use diaries use a different response format than many other similar instruments (i.e. the use of scanned responses rather than coding of text responses) and this may have an effect on the quality of the diary data and on which cases should be excluded. Cases considered to be of poor quality were removed from the main diary dataset and placed in a separate file so that they could be re-included for any analysis where the user thought they might be valuable.

2.5.1 Criteria for exclusion

Three criteria were used to exclude cases from the dataset.

Cases with large amounts of missing data

As mentioned above, it is common practice to remove cases with more than 90 minutes missing activity data from analyses. However, analyses of the LSAC data suggested that using this rule of thumb might be inappropriate as children who spent time away from their parents (e.g. in child care) were more likely to have greater levels of missing activity data. Instead, a diary was deleted from the file if it had no data of any kind for more than 90 minutes (or six time intervals). In Wave 1, this criterion excluded 239 diaries (3%) from the B-cohort file and 368 diaries (5%) from the K-cohort file. For Wave 2, 235 (4%) diaries were deleted from the B-cohort file and 268 (4%) were deleted from the K-cohort file. For Wave 3, 147 (3%) diaries were deleted from the B-cohort file and 233 (4%) were deleted from the K-cohort file.

Cases with large numbers of simultaneous activities

Most time use diaries request respondents to describe their main activity for each time period, with limited opportunities to describe secondary activities. The format of the LSAC time use diary meant that a number of activities could be specified separately; however, where numbers were large, it often indicated that the respondent had trouble understanding the task. As such, it was decided to exclude any respondent that gave more than five simultaneous activities for more than six time periods. In Wave 1, this criterion excluded 78 diaries (1%) from the B-cohort file and 55 diaries (1%) from the K-cohort file, while in Wave 2, 26 (0.4%) B-cohort diaries and 16 (0.2%) K-cohort diaries were deleted. For Wave 3, 11 (0.2%) diaries were deleted from the B-cohort file and 11 (0.2%) were deleted from the K-cohort file.

Cases with few changes in activities

Diaries with few changes in activities tended to occur when the parent either didn't have a good idea of the child's activity (e.g. large amount of time in non-parental care), or was not able to fill in the diary in detail. It was decided that fewer than 10 different activities over the 24-hour period represented an unacceptable lack of detail. This excluded 59 diaries (1%) from the B-cohort file and 144 diaries (2%) from K cohort file. In Wave 2, these figures were 110 (2%) and 171 (3%) respectively. For Wave 3, 120 (2%) diaries were deleted from the B-cohort file and 159 (3%) were deleted from the K-cohort file.

There were some diaries excluded for more than one reason, so in total for Wave 1, 330 diaries (4%) were excluded from the B-cohort file and 490 (7%) were excluded from the K-cohort file. The effect of the exclusion of these diaries on the socio-demographic composition of the time use diary sample can be seen in Table 5. The deleted diaries tended to come from lower socio-economic status families.

Table 5: Effect of deleting problem cases on socio-demographic composition of the Wave 1 LSAC TUD sample (unweighted data)

Wave 1	B cohort			K cohort		
	Full LSAC sample (%)	Full TUD sample (%)	Reduced TUD sample (%)	Full LSAC sample (%)	Full TUD sample (%)	Reduced TUD sample (%)
Gender						
Male	51.2	51.4	51.6	50.9	51.2	51.6
Female	48.8	48.6	48.4	49.1	48.8	48.4
Age range of children						
3-5 months / 51-53 months	11.2	11.1	11.2	10.6	10.6	10.7
6-11 months / 54-59 months	73.2	73.3	73.4	72.1	72.1	72.8
12-14 months / 60 62 months	14.7	14.8	14.8	16.1	16.1	15.7
15 19 months / 63–67 months	1.0	0.8	0.6	1.3	1.3	0.8
Family type						
Couple family:	90.7	91.5	93.0	86.0	87.0	88.9
both biological	90.1	91.0	92.5	82.9	84.3	86.6
other (e.g. step/blended)	0.6	0.6	0.5	3.1	2.8	2.3
Single parent family	9.3	8.5	7.0	14.0	12.9	11.0
Siblings						
Only child	39.5	40.0	40.7	11.5	11.2	10.8
One sibling	36.8	36.8	36.9	48.4	49.4	50.7
Two or more siblings	23.7	23.2	22.3	40.1	39.4	38.9
Cultural background						
Aboriginal or Torres Strait Islander	4.5	3.9	2.7	3.8	3.2	2.4
Mother speaks a language other than English at home	14.5	13.4	11.2	15.7	14.7	12.3
Work status						
Both parents or lone parent work/s	47.9	48.8	50.5	55.5	56.0	57.2
One parent works (in couple family)	40.8	41.2	41.5	32.8	33.5	34.2
No parent works	11.3	10.1	8.0	11.6	10.5	8.6
Educational status						
Mother completed Year 12	66.9	68.7	71.9	58.6	60.2	63.0
Father completed Year 12	58.5	59.4	60.8	52.7	53.3	54.6
Child care						
Child has a regular care arrangement (including school)	35.9	35.7	35.4	96.7	97.1	97.6
State						
New South Wales	31.6	30.8	30.0	31.6	31.2	30.8
Victoria	24.5	24.5	24.5	25.0	25.0	24.9
Queensland	20.6	20.9	21.2	19.8	20.1	20.4
South Australia	6.8	6.6	6.4	6.8	6.6	6.2
Western Australia	10.4	10.8	11.2	10.2	10.4	10.8
Tasmania	2.2	2.3	2.4	2.7	2.8	2.9
Northern Territory	1.7	1.8	1.8	1.7	1.6	1.5
Australian Capital Territory	2.1	2.3	2.4	2.3	2.3	2.4
Region						
Capital city statistical division	62.5	62.8	63.1	62.1	62.0	61.6
Balance of state	37.5	37.2	36.9	37.9	38.0	38.4
Number of observations (*N*) [a]	5,107	8,858	7,452	4,983	8,565	6,959

Note: [a]TUD samples are larger than the LSAC sample as respondents were asked to complete two diaries.

In Wave 2, 335 (5%) were deleted from the B-cohort file and 405 (6%) from the K-cohort file. The effect of the exclusion of these diaries on the socio-demographic composition of the time use diary sample can be seen in Table 6. Again, the deleted diaries tended to come from lower socio-economic status families.

Table 6: Effect of deleting problem cases on socio-demographic composition of the Wave 2 LSAC TUD sample (unweighted data)

	B cohort			K cohort		
	Full LSAC sample (%)	Full TUD sample (%)	Reduced TUD sample (%)	Full LSAC sample (%)	Full TUD sample (%)	Reduced TUD sample (%)
Gender						
Male	51.1	51.3	51.2	51.0	51.7	51.9
Female	48.9	48.7	48.8	49.0	48.3	48.2
Age range of children						
27- 32 months / 75-77 months	6.3	6.6	6.7	7.1	7.5	7.6
30-35 months / 78-83 months	64.8	66.1	66.5	63.7	64.7	64.9
36-38 months / 84-86 months	23.5	23.0	22.8	23.8	23.3	23.2
39-43 months / 87-91 months	5.4	4.3	3.9	5.4	4.5	4.3
Family type						
Couple family:	89.0	91.6	92.0	85.2	88.2	88.9
both biological	88.0	90.5	91.0	81.3	85.2	85.9
other (e.g. step/blended)	1.0	1.1	1.1	3.9	3.1	2.9
Single parent family	11.0	8.4	8.0	14.8	11.8	11.1
Siblings						
Only child	19.3	19.1	18.9	9.1	8.7	8.8
One sibling	49.1	51.4	51.9	45.2	47.7	48.1
Two or more siblings	31.6	29.6	29.2	45.7	43.6	43.1
Cultural background						
Aboriginal or Torres Strait Islander	3.9	2.5	2.3	3.4	2.3	2.3
Mother speaks a language other than English at home	13.4	11.8	11.1	14.7	13.6	12.5
Work status						
Both parents or lone parent work/s	56.9	58.0	58.5	65.4	67.6	68.3
One parent works (in couple family)	33.8	35.0	35.0	26.1	26.0	25.8
No parent works	9.3	7.0	6.5	8.6	6.5	5.9
Educational status						
Mother completed Year 12	69.0	73.0	74.1	60.1	63.8	64.9
Father completed Year 12	59.7	62.2	62.6	53.2	55.5	56.0
Child care						
Child has a regular care arrangement (including school)	70.4	71.3	71.5	99.7	99.7	99.6
State						
New South Wales	31.1	30.9	30.9	31.1	30.9	31.0
Victoria	24.3	24.8	24.7	24.3	24.5	24.2
Queensland	21.4	21.1	21.1	21.4	20.4	20.8
South Australia	6.7	6.8	6.8	6.7	6.9	6.8
Western Australia	10.6	10.4	10.5	10.6	10.4	10.4
Tasmania	2.3	2.3	2.4	2.3	3.2	3.2
Northern Territory	1.4	1.3	1.3	1.4	1.3	1.2
Australian Capital Territory	2.2	2.3	2.3	2.2	2.4	2.4
Region						
Capital city statistical division	61.9	62.6	62.5	61.6	61.4	61.4
Balance of state	38.1	37.4	37.5	38.4	38.6	38.6
Number of observations (N)[a]	4,606	6,917	6,582	4,464	6,858	6,483

Note: [a] TUD samples are larger than the LSAC sample as respondents were asked to complete two diaries.

In Wave 3, 228 (4%) were deleted from the B-cohort file and 339 (6%) from the K-cohort file. The effect of the exclusion of these diaries on the socio-demographic composition of the time use diary sample can be seen in Table 7. Again, the deleted diaries tended to come from lower socio-economic status families.

Table 7: Effect of deleting problem cases on socio-demographic composition of the Wave 3 LSAC TUD sample (unweighted data)

	B cohort			K cohort		
	Full LSAC sample (%)	Full TUD sample (%)	Reduced TUD sample (%)	Full LSAC sample (%)	Full TUD sample (%)	Reduced TUD sample (%)
Gender						
Male	51.3	52.1	52.1	51.1	51.4	51.2
Female	48.7	47.9	47.9	48.9	48.6	48.8
Age range of children						
27-32 months / 75-77 months	7.8	8.4	8.5	8.4	8.8	8.8
30-35 months / 78-83 months	67.2	68.5	68.5	65.7	65.1	65.1
36-38 months / 84-86 months	20.7	19.9	19.8	21.9	22.5	22.5
39-43 months / 87-91 months	4.3	3.1	3.1	4.1	3.7	3.6
Family type						
Couple family:	88.9	91.6	91.8	85.6	86.4	86.7
both biological	85.8	89.0	89.3	78.8	79.7	80.2
other (e.g. step/blended)	3.0	2.6	2.3	6.8	6.7	6.5
Single parent family	11.1	8.4	8.2	14.4	13.6	13.3
Siblings						
Only child	10.4	10.4	10.2	8.2	8.6	8.6
One sibling	48.1	51.2	51.6	44.1	45.2	45.2
Two or more siblings	41.5	38.4	38.2	47.7	46.2	46.2
Cultural background						
Aboriginal or Torres Strait Islander	3.4	2.0	2.0	2.9	2.6	2.4
Mother speaks a language other than English at home	12.6	10.7	10.2	13.8	13.5	13.2
Work status						
Both parents or lone parent work/s	63.0	64.0	64.2	72.8	73.1	73.4
One parent works (in couple family)	29.7	30.9	30.9	20.7	20.7	20.7
No parent works	7.4	5.1	4.9	6.5	6.2	5.9
Educational status						
Mother completed Year 12	69.8	74.1	74.8	61.3	62.2	62.8
Father completed Year 12	60.4	63.3	63.7	54.0	54.9	55.1
Child care						
Child has a regular care arrangement (including school)	96.6	97.4	97.4	99.5	99.4	99.4
State						
New South Wales	30.1	29.6	29.6	30.8	31.2	31.5
Victoria	24.6	24.3	24.0	24.4	19.6	19.5
Queensland	22.0	21.4	21.5	20.8	23.2	23.1
South Australia	7.0	7.6	7.7	6.9	5.1	5.1
Western Australia	10.3	10.7	10.7	10.2	14.9	14.9
Tasmania	2.4	2.6	2.6	30.	2.4	2.4
Northern Territory	1.2	1.3	1.3	1.4	1.6	1.5
Australian Capital Territory	2.4	2.6	2.6	2.5	2.0	2.1
Region						
Capital city statistical division	61.9	62.6	62.6	61.3	61.3	61.0
Balance of state	38.2	37.4	37.4	38.7	38.8	39.0
Number of observations (*N*)[a]	4,384	5,909	6,582	4,332	5,924	5,585

Note: [a] TUD samples are larger than the LSAC sample as respondents were asked to complete two diaries.

2.6 Summary

Corrections to improve data quality and deletion of problem cases had some effect on the rate of false positives due to scanning errors in the corrected files (i.e. the cases that had been checked against the paper forms). When these improvements were performed on the corrected file, the false positive rate dropped to 5% for the K-cohort file and 3% for the B-cohort file. In Wave 2, these same recodes applied to more rigorously checked, scanned files recoded far fewer responses, which adds further evidence that it was largely false positive responses that were being recoded in Wave 1. Tables 8a to 8f show the effect of the recodes on estimates produced by the full data file. In the final Wave 1 B-cohort file (i.e. with cases deleted and all corrections made) 88% of cases had at least one correction made to them, while in the final K-cohort file, 84% of cases had at least one correction. In Wave 2, these proportions were much lower: 41% for the B cohort and 48% for the Kcohort and, in Wave 3, they were 42% for the B cohort and 46% for the K cohort.

Table 8a: Effect of recoding and cases deletions on estimates of time use in number of minutes/day (B cohort, Wave 1)[a]

	Raw file	After recodes	After recodes and deletions
What was the child doing?			
Not sure what child was doing	42.5	19.8	19.0
Sleeping, napping	772.5	784.5	800.3
Awake in bed/cot	48.2	46.9	45.2
Looking around, doing nothing	28.2	27.6	25.8
Bathing/nappy change, dress/hair care	91.5	91.3	90.9
Breastfeeding	53.7	52.9	52.3
Other eating, drinking, being fed	129.4	128.9	129.1
Crying, upset	43.4	42.9	42.1
Destroying things, creating mess	22.6	22.3	21.1
Held, cuddled, comforted, soothed	129.8	127.5	128.0
Watching TV, video or DVD	37.3	37.3	36.0
Listening to tapes, CDs, radio, music	29.3	28.4	26.9
Read a story, talked/sung to, sing/talk	77.9	77.1	76.2
Colour/drawing, look at book, puzzles	8.5	8.5	7.8
Organised activities/play group	10.6	10.5	9.6
Crawl, climb, swing arms or legs	121.7	120.2	119.2
Other play, other activities	137.4	135.7	137.9
Visiting people, special event, party	41.2	40.3	39.1
Taken places with adult (e.g. shopping)	54.7	54.1	53.2
Travel			
Taken out in pram or bicycle seat	36.8	36.4	35.3
Travel in car/other household vehicle	54.9	54.4	54.2
Travel on public transport, ferry, plane	3.0	2.9	2.1
Where was the child?			
Own home (indoors)	1111.0	1100.9	1130.0
Other person's home (indoors)	68.4	68.4	68.7
Day care centre/play group	22.2	20.9	19.7
Other indoors	43.7	43.7	43.6
Other outdoors	66.3	66.3	67.2
In the same room, nearby if outside			
Alone	370.5	366.4	380.2
Mother, step-mother	781.6	756.3	775.7
Father, step-father	405.2	393.9	403.7
Grandparent(s)/other adult relative	105.0	103.0	102.9
Brother(s), sister(s), other children	365.5	364.4	371.5
Other adult(s)	65.2	71.8	71.7
Dog, cat or other pet (not fish)	119.4	114.3	116.3
Payment			
Someone paid for this activity	21.9	21.9	21.4

Note: [a] Analysis uses weights that adjust for general LSAC non-response as well as weighting each day of the week equally.in the analysis. These weights are recalculated when the poor quality cases are deleted.

Table 8b: Effect of recoding and cases deletions on estimates of time use in number of minutes/day (K cohort, Wave 1)[a]

	Raw file	After recodes	After recodes and deletions
What was the child doing?			
Not sure what child was doing	68.4	39.0	38.4
Sleeping, napping	626.6	634.7	650.8
Awake in bed	37.3	33.8	34.0
Eating, drinking, being fed	124.0	120.7	123.0
Bathing, dressing, hair care, health care	59.9	58.8	59.6
Do nothing, bored/restless	7.7	7.0	6.5
Crying, upset, tantrum	10.1	9.1	8.9
Destroy things, create mess	8.0	6.9	6.3
Held, cuddled, comforted, soothed	40.5	37.5	37.0
Being reprimanded, corrected	12.3	12.0	11.4
Watching TV, video, DVD, movie	124.8	121.1	122.6
Listening to tapes, CDs, radio, music	18.9	17.9	17.5
Use computer/computer games	16.6	15.9	15.5
Read a story, talk/sing, talked/sung to	61.3	59.7	59.8
Colour, look at book, educational game	43.0	42.2	41.7
Being taught to do chores, read, etc.	18.3	17.5	17.1
Walk for travel or for fun	13.7	13.1	12.8
Ride bicycle, trike etc. (travel or fun)	17.9	17.1	16.5
Other exercise—swim /dance/run about	47.4	45.7	46.4
Visiting people, special event, party	50.4	46.9	47.2
Other play, other activities	109.6	105.1	108.1
Travel in pusher or on bicycle seat	3.9	3.4	3.0
Travel in car/other household vehicle	57.6	55.4	56.6
Travel on public transport, ferry, plane	5.5	4.7	4.3
Taken places with adult (e.g. shopping)	49.5	47.4	47.9
Organised lessons/activities	73.8	71.9	74.5
Where was the child?			
Own home (indoors)	917.4	906.5	941.2
Other person's home (indoors)	65.4	64.9	64.7
Day care centre/playgroup	108.6	104.9	106.3
Other indoors	73.0	73.0	75.6
Other outdoors	107.3	102.2	105.2
In the same room, nearby if outside			
Alone	226.4	225.8	236.9
Mother, step-mother	593.4	593.4	616.3
Father, step-father	343.5	343.5	357.7
Grandparent(s)/other adult relative	92.8	92.8	92.7
Brother(s), sister(s), other children	662.8	711.0	738.8
Other adult(s)	123.8	172.0	175.5
Dog, cat or other pet (not fish)	127.8	127.8	131.7
Payment			
Someone paid for this activity	74.1	74.1	77.2

Note: [a] Analysis uses weights that adjust for general LSAC non-response as well as weighting each day of the week equally in the analysis. These weights are recalculated when the poor quality cases are deleted.

Table 8c: Effect of recoding and cases deletions on estimates of time use in number of minutes/day (B cohort, Wave 2)[a]

	Raw file	After recodes	After recodes and deletions
What was the child doing?			
Not sure what child was doing	60.2	48.6	47.4
Sleeping, napping	662.3	673.4	686.6
Awake in bed	40.7	41.0	41.0
Eating, drinking, being fed	117.7	119.1	120.6
Bathing, dressing, hair care, health care	54.0	54.7	55.2
Doing nothing, bored/restless	5.9	6.0	5.7
Crying, upset, tantrum	12.1	12.3	12.3
Arguing, fighting	5.6	5.7	5.4
Destroy things, create mess	7.1	7.2	6.5
Being reprimanded	9.7	9.8	9.1
Being held, cuddled, comforted, soothed	45.8	46.2	45.8
Watching TV, video, DVD, movie	94.5	95.1	95.9
Listening to tapes, CDs, radio, music	20.4	20.5	20.6
Using computer, computer game	4.3	4.3	4.1
Read a story, told a story, sung to	33.6	33.9	34.3
Colour/draw, look at book, educational game	36.2	36.5	36.3
Quiet free play	76.6	76.9	78.7
Active free play	88.8	89.1	90.4
Being taught to do chores	11.7	11.7	11.4
Visiting people, special event, party	72.0	72.1	73.4
Organised lessons/activities	15.8	15.8	16.0
Travel			
Walking	13.0	13.1	12.5
Ride bicycle/trike	9.4	9.5	9.0
Travel in car	51.7	52.0	52.0
Travel in a pusher/bicycle seat	5.3	5.3	5.1
Travel on public transport	1.5	1.5	1.3
Taken places with adult (e.g. shopping)	34.5	34.6	34.7
Where was the child?			
Own home (indoors)	944.8	944.8	974.8
Other person's home (indoors)	61.2	61.2	61.7
Day care centre/play group	87.5	87.5	85.4
Other indoors	85.5	85.5	86.9
Other outdoors	68.6	68.6	70.4
In the same room, nearby if outside			
Alone	298.3	298.3	311.2
Mother, step-mother	677.7	677.7	699.3
Father, step-father	368.2	368.2	379.2
Grandparent(s)/other adult relative	94.0	94.0	94.9
Brother(s), sister(s), other children	551.3	590.0	606.4
Other adult(s)	91.1	129.8	129.0
Dog, cat or other pet (not fish)	113.9	113.9	118.3
Payment			
Someone paid for this activity	53.3	53.3	54.3

Note: [a] Analysis uses weights that adjust for general LSAC non-response as well as weighting each day of the week equally in the analysis. These weights are recalculated when the poor quality cases are deleted.

Table 8d: Effect of recoding and cases deletions on estimates of time use in number of minutes/day (K cohort, Wave 2)[a]

	Raw file	After recodes	After recodes and deletions
What was the child doing?			
Not sure what child was doing	99.2	86.6	86.1
Sleeping, napping	598.4	607.5	620.2
Awake in bed	30.8	31.1	30.5
Eating and drinking	95.8	96.9	98.7
Bathing, dressing, hair care, health care	49.6	50.2	51.0
Do nothing, bored/restless	4.1	4.1	3.7
Crying, upset, tantrum	3.0	3.0	2.8
Arguing, fighting, destroy things	3.8	3.8	3.5
Held, cuddled, comforted, soothed	18.9	19.1	18.3
Being reprimanded, corrected	6.8	6.9	6.6
Watching TV, video, DVD, movie	91.2	91.7	92.9
Listening to tapes, CDs, radio, music	12.3	12.4	12.2
Use computer/computer games	18.4	18.6	18.5
Read a story, talk/sing, talked/sung to	16.9	17.1	17.1
Reading looking at book by self	21.0	21.2	21.3
Quiet free play	47.8	47.9	49.5
Active free play	65.6	65.8	67.6
Helping with chores/jobs	19.0	19.2	19.0
Visiting people, special event, party	59.7	59.8	59.9
Organised sport/physical activity	18.7	18.8	18.8
Other organised lessons/activities	17.3	17.4	18.0
Travel			
Walk for travel or for fun	9.6	9.6	9.4
Ride bicycle, trike, etc. (travel or fun)	11.4	11.5	11.5
Travel in car	47.8	48.0	49.0
Travel on public transport	4.7	4.7	4.7
Taken places with adult (e.g. shopping)	20.4	20.5	20.7
Where was the child?			
Own home (indoors)	784.2	784.2	812.8
Own home outdoors	56.7	56.7	57.0
School, after/before school care	219.8	219.8	223.1
Other indoors	78.1	78.1	78.1
Other outdoors	65.5	65.5	67.3
In the same room, nearby if outside			
Alone	247.4	247.4	259.3
Mother, step-mother	479.1	479.1	494.1
Father, step-father	303.3	303.3	312.7
Grandparent(s)/other adult relative	65.5	65.5	65.4
Brother(s), sister(s), other children	649.3	759.6	781.9
Other adult(s)	141.6	251.9	257.2
Dog, cat or other pet (not fish)	126.2	126.2	130.2
Homework			
Activity done as part of homework	11.7	11.7	11.5

Note: [a] Analysis uses weights that adjust for general LSAC non-response as well as weighting each day of the week equally in the analysis. These weights are recalculated when the poor quality cases are deleted.

Table 8e: Effect of recoding and cases deletions on estimates of time use in number of minutes/day (B cohort, Wave 3)[a]

	Raw file	After recodes	After recodes and deletions
What was the child doing?			
Not sure what child was doing	60.0	48.5	47.7
Sleeping, napping	626.8	635.2	646.9
Awake in bed	35.9	36.2	35.9
Eating, drinking, being fed	106.5	107.7	109.5
Bathing, dressing, hair care, health care	53.4	54.0	55.0
Doing nothing, bored/restless	3.7	3.8	3.6
Crying, upset, tantrum	4.8	4.9	4.9
Arguing, fighting	5.7	5.8	5.7
Destroy things, create mess	3.3	3.3	3.2
Being reprimanded, corrected	6.6	6.7	6.6
Being held, cuddled, comforted, soothed	28.4	28.7	29.3
Watching TV, video, DVD, movie	99.1	99.7	100.9
Listening to tapes, CDs, radio, music	15.9	16.0	15.9
Using computer, computer game	14.0	14.1	14.3
Read a story, told a story, sung to	45.9	46.3	47.2
Colour/draw, look at book, educational game	39.4	39.6	39.9
Quiet free play	65.5	65.8	67.3
Active free play	82.0	82.3	83.8
Being taught to do chores	15.3	15.4	15.4
Visiting people, special event, party	63.5	63.5	64.2
Organised lessons/activities	54.5	54.5	55.6
Travel			
Walking	9.7	9.8	9.7
Travel in a pusher/bicycle seat	2.4	2.4	2.3
Travel in car	49.5	49.8	50.7
Travel on public transport	3.3	3.3	3.3
Taken places with adult (e.g. shopping)	24.8	24.8	25.1
Ride bicycle/trike, etc.	8.1	8.1	8.1
Where was the child?			
Own home (indoors)	906.1	906.1	929.9
Other person's home (indoors)	57.8	57.8	57.3
Day care centre/play group/pre-school/school	151.9	151.9	152.7
Other indoors	43.4	43.4	44.4
Other outdoors	86.8	86.8	89.2
In the same room, nearby if outside			
Alone	212.7	212.7	219.0
Mother, step-mother	690.5	690.5	707.8
Father, step-father	412.6	412.6	422.6
Grandparent(s)/other adult relative	85.4	85.4	87.1
Brother(s), sister(s), other children	719.0	778.1	798.9
Other adult(s)	132.6	191.7	194.0
Dog, cat or other pet (not fish)	160.3	160.3	165.0
Payment			
Someone paid for this activity	67.3	67.3	67.8

Note: [a] Analysis uses weights that adjust for general LSAC non-response as well as weighting each day of the week equally in the analysis. These weights are recalculated when the poor quality cases are deleted.

Table 8f: Effect of recoding and cases deletions on estimates of time use in number of minutes/day (K cohort, Wave 3)[a]

	Raw file	After recodes	After recodes and deletions
What was the child doing?			
Not sure what child was doing	91.2	75.4	75.1
Sleeping, napping	584.1	592.4	606.6
Awake in bed	34.6	34.8	34.2
Eating and drinking	95.8	96.7	97.8
Bathing, dressing, hair care, health care	48.8	49.4	50.0
Do nothing, bored/restless	5.1	5.2	4.3
Sulking, upset	3.0	3.0	2.5
Arguing, fighting	5.2	5.3	4.8
Being hugged, comforted, etc.	11.6	11.7	11.2
Being reprimanded, corrected	6.8	6.9	6.2
Watching TV, video, DVD, movie	103.5	104.1	105.0
Listening to tapes, CDs, radio, music	11.9	12.0	11.3
Use computer/computer games	30.2	30.4	30.2
Read a story, talk/sing, talked/sung to	9.9	10.0	9.4
Reading /looking at book by self	24.7	24.9	25.0
Quiet free play	42.6	42.8	43.7
Active free play	62.6	62.8	64.0
Helping with chores/jobs	22.8	22.9	22.8
Visiting people, special event, party	60.6	60.7	60.9
Organised sport/physical activity	25.4	25.4	25.9
Other organised lessons / activities	22.5	22.6	23.0
Travel			
Walk for travel or for fun	10.1	10.2	9.8
Ride bicycle, trike, etc. (travel or fun)	12.0	12.0	11.3
Travel in car	49.0	49.3	49.2
Travel on public transport	5.5	5.5	5.5
Taken places with adult (e.g. shopping)	20.9	21.0	20.6
Where was the child?			
Own home (indoors)	809.9	809.9	838.1
Own home outdoors	53.6	53.6	53.4
School, after/before school care	209.2	209.2	215.5
Other indoors	91.3	91.3	92.6
Other outdoors	65.7	65.7	67.4
In the same room, nearby if outside			
Alone	247.3	247.3	260.6
Mother, step-mother	524.3	524.3	542.9
Father, step-father	344.4	344.4	355.8
Grandparent(s)/other adult relative	65.1	65.1	66.0
Brother(s), sister(s)	599.6	599.6	622.0
Other children	219.1	306.1	316.6
Other adult(s)	161.8	248.8	256.7
Dog, cat or other pet (not fish)	165.5	165.5	172.6
Homework			
Activity done as part of homework	14.1	14.1	13.7

Note: [a] Analysis uses weights that adjust for general LSAC non-response as well as weighting each day of the week equally in the analysis. These weights are recalculated when the poor quality cases are deleted.

Acknowledgement

This chapter is largely based on the work of Jude Brown and Michael Bittman of the University of New England. David Zago of the Australian Bureau of Statistics provided the information on the process used to scan Waves 2 and 3 forms.

3 Report on Adapted PPVT-III and "Who Am I?"

Sheldon Rothman

Australian Council for Educational Research

3.1 Wave 1 scoring

The first wave of the Longitudinal Study of Australian Children (LSAC) used two tests with the four-year-old sample. The Adapted PPVT-III is a shortened version of the *Peabody Picture Vocabulary Test,* Third Edition (Dunn & Dunn, 1997), which is a test of receptive vocabulary used as a screening test of verbal ability. This adaptation is based on work done in the USA for the Head Start Impact Study, with a number of changes for use in Australia. "Who Am I?" (de Lemos & Doig, 2000) assesses the cognitive processes that underlie the learning of early literacy and numeracy skills. One item was added to the standard "Who Am I?" booklet for use in LSAC. Summary statistics for each test are shown in Table 9.

Table 9: Summary statistics for administration of the adapted PPVT-III and "Who Am I?" tests as part of LSAC Wave 1

	Adapted PPVT	Who Am I?
Number of cases	4407	4827
Mean scaled scores	64.2 (se = 0.123)	63.8 (se = 0.125)
Mean number of items correct/mean raw score	28.2 (se = 0.086)	25.6 (se = 0.103)
Minimum number of items correct	2	0
Maximum number of items correct	40	44
Reliability	0.76	0.89

Note: For the adapted PPVT-III, it was assumed that children who were not required to answer 10 "basal" items had answered these items correctly. Reliability reported here is the person separation reliability (Wright & Masters, 1982).

3.1.1 Adapted PPVT-III

The PPVT-III was adapted for use in LSAC by altering the administration procedures, reducing the number of items administered during testing. To determine which items to retain for the adapted version, 215 children aged from 41 to 66 months (mean = 54.7 months) were given the PPVT-III, with test administrators following standard procedures. After testing, a one-parameter (Rasch) item response model was fitted to the data, which consisted of correct and incorrect responses. The person separation reliability was 0.88. After determining the "best" 40 items for use in a shortened version, the remaining items were then fit again to a one-parameter item response model; the person separation reliability decreased to 0.78.

Development of the model suggested that 37% of children would require only the core set of items, 5% would require the core and basal sets, and 58% would require the core and ceiling sets, resulting in an average of 26.3 items administered. The Pearson product-moment correlation between the full PPVT-III and the adapted PPVT-III was 0.93 for all children, and 0.91 for 4 year olds (Rothman, 2003).

3.1.2 Scaling

The adapted PPVT-III was scaled using a two-stage process. In the first stage, only the core set of 20 items was used, as these items had been administered to all children. For these core items, Rasch estimates were determined for each item, providing an indication of their difficulty. In the second stage, all 40 items were fitted, using the item estimates for the core items as anchors. This gave item estimates for the basal and ceiling items relative to the core items. The final case estimates were then transformed to a scale with a mean of 64 and standard deviation of 8.

3.1.3 Results

In Wave 1, 4,407 children were administered the adapted PPVT-III. Children ranged in age from 43 months to 79 months (mean = 57.3); 21% were aged 60 months or older. Twenty-one per cent of children were administered only the core set, 1% were given the core and basal sets, and 78% were given the core and ceiling sets, resulting in an average of 27.9 items. The test had a person separation reliability of 0.76.

3.1.4 Quality of the PPVT test

The statistics indicate that the core and ceiling items used for the adapted PPVT-III test fit the Rasch model well. This is shown in Figure 3, the item fit map. The infit mean square ranged from 0.86 to 1.17 for items 11-30 (the core set) and items 31-40 (the ceiling set). On each of the items in the basal set (items 1-10), the infit mean square was extremely low (0.49 or less) because only 30 children (1%) were administered these items; all other children were assumed to have correctly answered these items. The item map in Figure 3, which shows the item estimates (difficulties) mapped against the case estimates (children's ability levels), shows that the basal items were appropriate for children given that set but that the core and ceiling items were relatively easy for those who were given those sets.

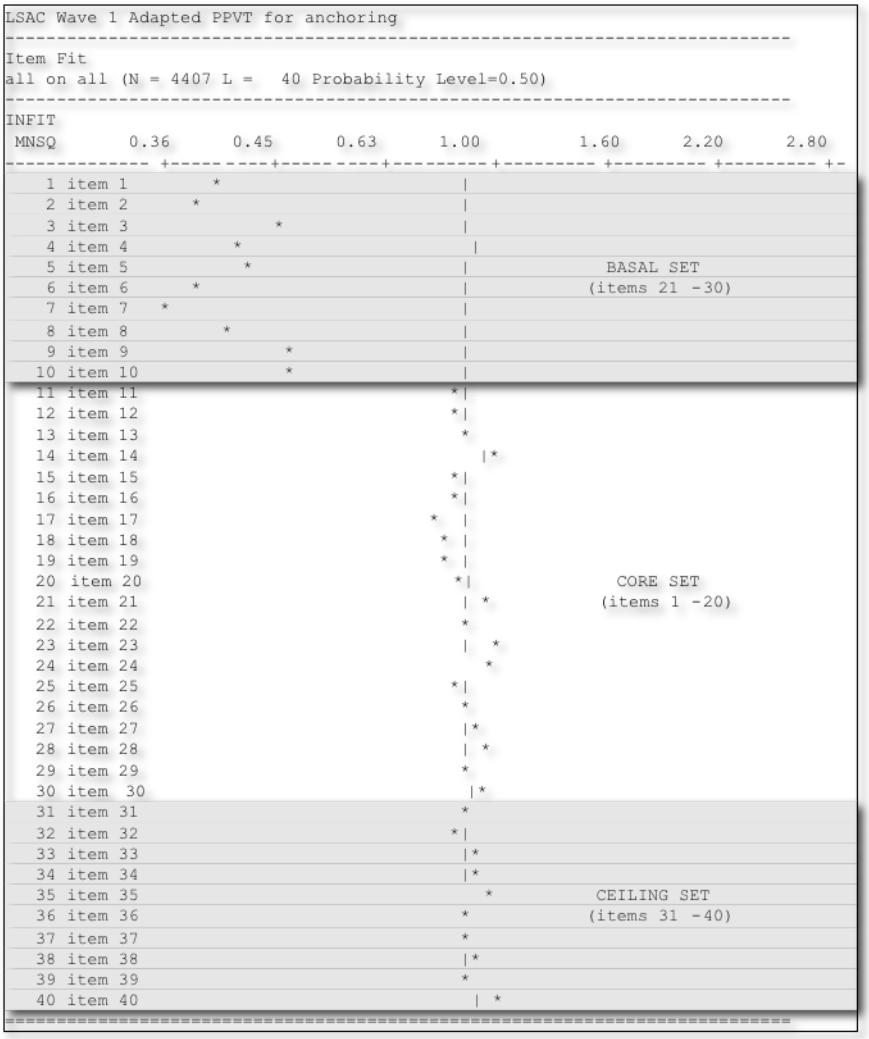

Figure 3: Item fit map for all items on the Australian adaptation of the Peabody Picture Vocabulary Test (PPVT-III) calibrated with all cases anchored to core items

3.1.5 "Who Am I?"

"Who Am I?" consisted of 11 pages on which children were to write their names, copy shapes and write words and numbers. Each response was assessed on a four-point scale relating to the skill required for the task. A score of zero was assigned if no attempt was made on the item. The data were fit using a partial credit item response model. The final case estimates were transformed to a scale with a mean of 64 and standard deviation of 8. Summary statistics are shown in Table 9 above.

In Wave 1, 4,827 children were administered "Who Am I?" The test had a person separation reliability of 0.89.

3.1.6 Quality of the "Who am I" test

The statistics indicate that the "Who Am I?" data fit the rating scale model well, with most items falling within acceptable ranges, as shown in Figure 4. The most difficult item on the test was item 10, in which children were asked to write a sentence. Only nine children received four points for their response; more than one-half of children made no attempt on this item. This is also shown in the item fit map (Figure 4): item 10 (Sentence) has an infit mean square of 1.67, while all other items ranged from 0.77 to 1.14.

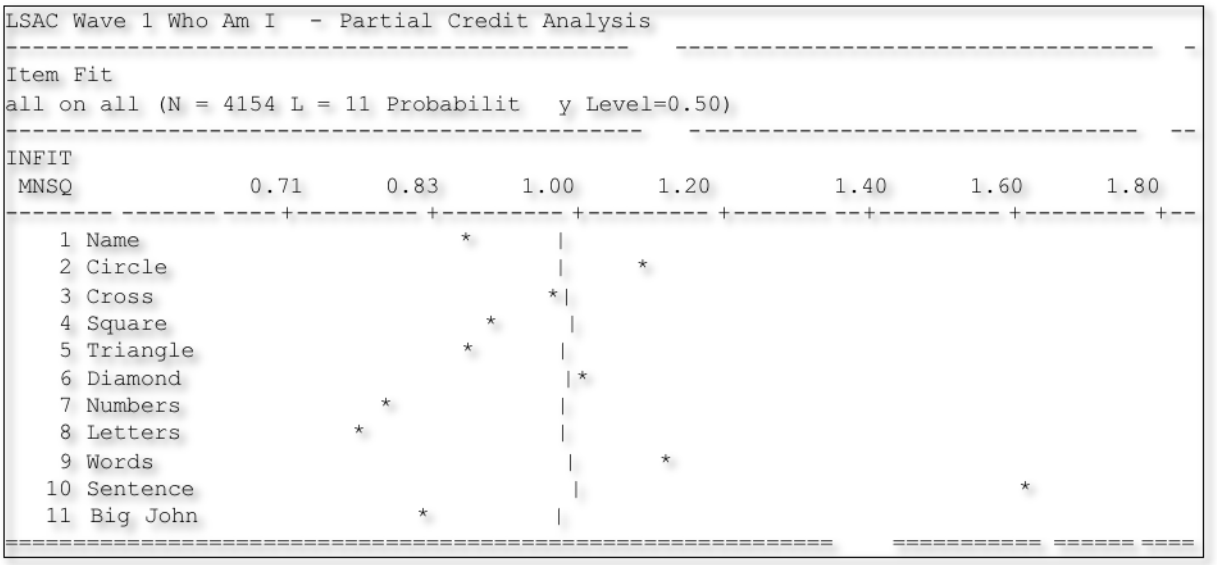

Figure 4: Item fit map for all items on the "Who Am I?" test

Analysis of a partial credit model provides information on the thresholds required to move from any score to a higher score on each item. This information is provided in Figure 5, the item map, and is plotted against the distribution of case estimates. For all items, higher scores had higher logit values, indicating that higher scores were achieved by children who had higher ability estimates. The item map also indicates that the distribution of children (on the left) was only slightly above the distribution of the items (on the right), indicating that there was a good match between the difficulty of the items and the children's ability levels.

3.1.7 Correlation between the adapted PPVT-III and "Who Am I?"

The two tests had a Pearson product-moment correlation of 0.309, based on 4,386 children who received scores on both tests. While significant at the .01 level, this is considered a low-to-moderate correlation.

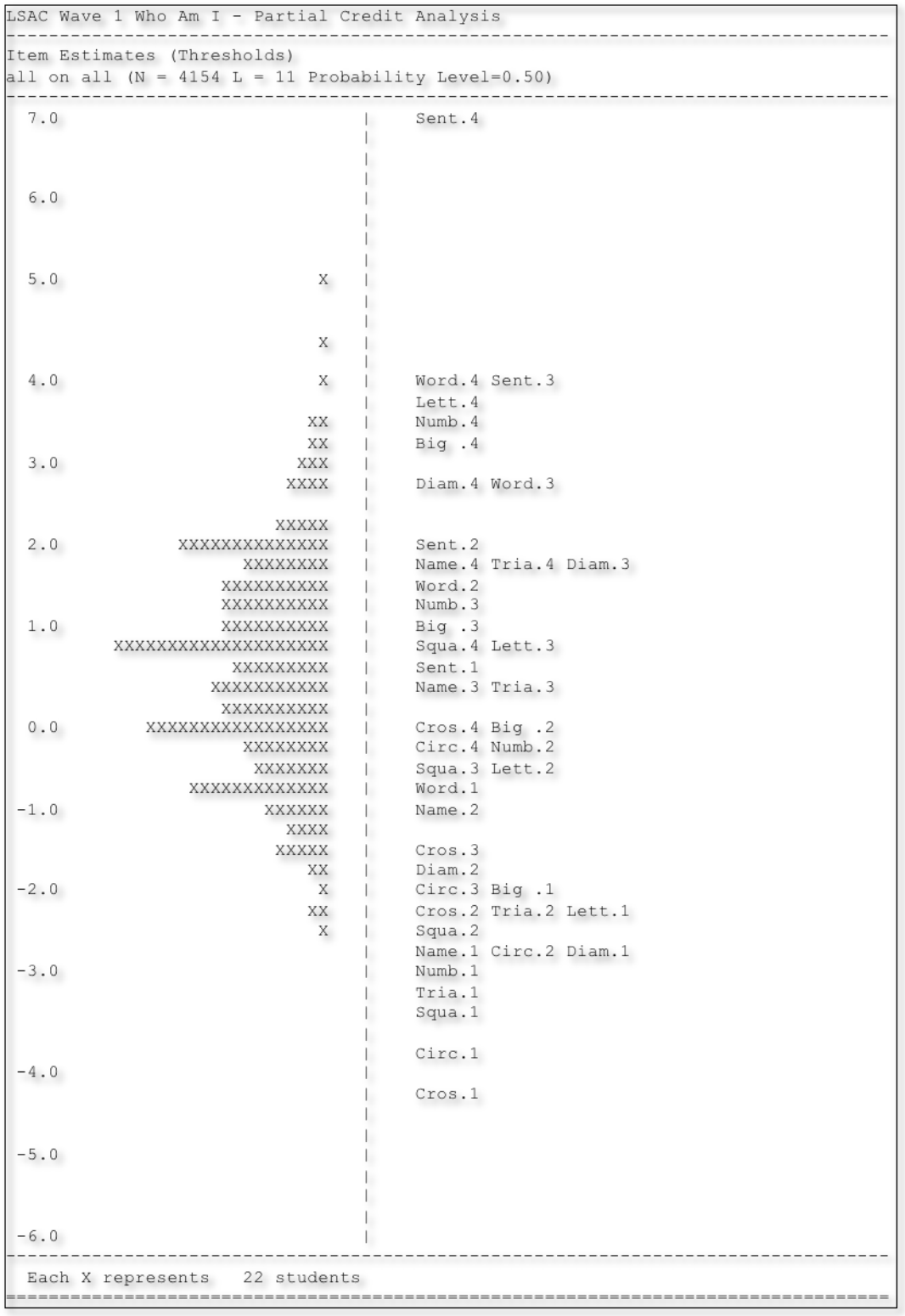

Figure 5: Item map for all cases on the “Who Am I?” test.

3.2 Wave 2 PPVT development

This paper describes the procedures used to develop a shortened version of the Peabody Picture Vocabulary Test (PPVT) for use in *Growing Up in Australia*: The Longitudinal Study of Australian Children (LSAC). This version of the test is to be used with 6 year olds as part of the second wave of LSAC and is linked to the shortened version developed in 2003 for use with 4 year olds as part of the first wave of LSAC (Rothman,

2003). The concept of a shortened version of the PPVT is from work done in the USA for the Head Start Impact Study. The original work was described in a draft paper by Philip Fletcher of Westat.

3.2.1 Procedures

All procedures described below are based on procedures used for the version used with 4 year olds. For that version, four alternative sets of items were tested; one set was developed for use in LSAC. For the 6-year-old version, no alternative sets were used as the scope of the project was to develop a test that could be linked to the 4-year-old version.

As done for the test administered to 4 year olds, the purpose was to develop a test that would consist of 40 items divided into a core set of 20 items, a basal set of 10 items for children who miss a minimum number of items on the core set, and a ceiling set of 10 items for children who correctly answer a minimum number of items on the core set. No child would take more than 30 items. It was also decided that at least 50% of children should be required to take the core set only.

3.2.2 Testing

A sample comprising 421 children was drawn from schools in New South Wales, Victoria and Queensland. During July and August 2005, the children were administered the full version of the PPVT-III, Form A, using the standard procedures for administering the test to 6 year olds. These children ranged in age from 5 years 7 months to 7 years 11 months. Seventy-eight per cent of the children were six years old, and 18% were 7 years old. All children were in the same classes at the schools involved in the data collection. Subsequent examination of the data showed that the children from out-of-range ages did not appear as significantly different cases.

3.2.3 Analysis

Test items were examined using a one-parameter logistic IRT model with the software Quest. For items below the PPVT basal set that were not administered, all were marked as correct. Items with a low number of responses were eliminated from the IRT analysis. Overall, 132 items were used for analysis, as they covered a range that would allow 40 items to be selected and included the items administered in the 4-year-old test.

3.2.4 Selection of items

The properties of the items were then determined, based on the data available from the Quest output. The first stage was to identify link items from the 4-year-old test that could be used with the 6 year olds. For the 20 items of the core set, eight items that had appeared in the four-year-old test were selected. These items were selected on the basis of infit mean square and outfit mean square close to 1.00 in both administrations, the degree of difficulty on the items among both groups, the consistency of change between the administration to the groups, and the ability to provide a reasonable spread across the core set. Two items from the 4-year-old test were selected for the basal set, and one item from the 4-year-old test was selected for the ceiling set.

After the link items were selected, the remaining items were selected using those with infit mean square and outfit mean square close to 1.00, good discrimination and an ability to provide a reasonable range of item difficulties (–2.50 to +2.50). Items were also selected according to their position in the original PPVT sets and their parts of speech: nouns, verbs and adjectives. The final 20 core items were then positioned into two sets of 10 items, with the first 10 items generally easier than the second 10 items but with an overlap of item estimates. Similar analyses were done to select the 10 basal and 10 ceiling items.

Table 10: Items selected for adaptive PPVT-III for use with 6 year olds in LSAC

Set	PPVT-III Form A item number	Item	Item threshold	Infit mean square
Core 1	42	harp*	-2.55	1.01
	74	nostril*	-2.29	0.96
	56	furry*	-2.08	0.96
	52	diving*	-1.99	1.02
	78	horrified*	-1.44	0.99
	67	calculator	-0.38	1.10
	77	towing	-0.12	1.02
	91	clarinet	-0.02	1.07
	107	fern	0.53	1.03
	118	archery	0.88	0.98
Core 2	66	swamp*	-0.47	1.13
	90	interviewing*	-0.20	1.00
	96	vine*	0.10	0.97
	88	surprised	0.61	1.02
	68	signal	0.91	1.03
	114	injecting	0.97	0.99
	128	wailing	1.29	0.94
	131	foundation	1.85	0.98
	140	pastry	2.33	0.99
	125	valve	2.74	0.98
Basal	45	juggling	-4.98	0.74
	32	fountain*	-3.85	0.97
	40	farm*	-3.26	0.99
	47	tearing	-2.98	0.77
	49	parachute	-2.19	0.93
	71	vegetable	-1.70	1.04
	57	drilling	-1.62	0.92
	61	vehicle	-1.30	0.99
	75	vase	-1.21	0.94
	85	flamingo	-0.52	0.97
Ceiling	122	dilapidated*	1.11	0.98
	97	pedal	1.85	1.03
	149	abrasive	1.97	1.09
	143	pedestrian	2.07	0.97
	117	microscope	2.15	1.07
	153	detonation	2.69	0.94
	151	cascade	2.96	0.91
	139	consuming	3.57	1.04
	148	replenishing	4.58	1.14
	167	talon	- -	- -

Notes: Item threshold and infit mean square statistics are from the simulated test. *Link item included in test for 4 year olds.

3.2.5 Simulations

Simulation results were then calculated in SPSS. With the objective of having approximately 50% of children requiring only the core set of 20 items, the items were checked to determine percentages of children requiring the basal or ceiling sets. Rules governing the administration of the test, particularly those regarding the number of correct items required for administration of the basal or ceiling sets, also guided the selection

of items. The simulation suggests that approximately 25% of children will require the ceiling set, 10% the basal set and 65% the core only. It should be noted that similar targets for the 4 year olds were not achieved in the first wave of LSAC, with more than 65% of children requiring the ceiling set.

Once the 40 items were selected, a new IRT analysis was conducted using only those items. Case and item estimates showed that the model fit the data well.

Results for the two versions—the full PPVT and the adapted version—were then compared: the full PPVT raw score with the case estimates from both the full test and the adapted test (Figure 6). The lowest correlation was 0.887; most correlations were in the 0.93 to 0.97 range, suggesting that the adapted version of the PPVT provides similar results to the full PPVT.

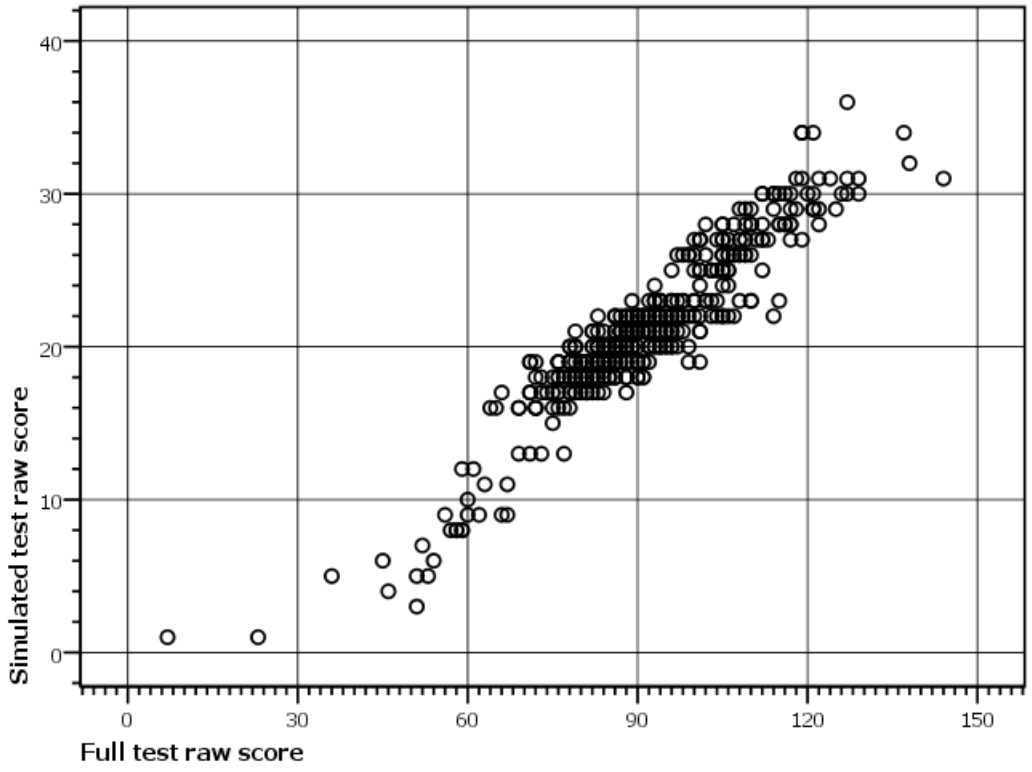

Figure 6: Scatterplot showing joint distribution of scores on simulated adaptive PPVT-III and scores on full PPVT-III for 6 year olds.

3.2.6 Comment

The 6-year-old version of the adapted PPVT-III for use in LSAC was developed so that it can be linked with the 4-year-old version of the test. This allows the measurement of growth between administrations of the test. Forty items were selected for the test, with 20 items administered to all children as the core set (core sets 1 and 2). For children who make 15 to 20 errors, an additional basal set of 10 items is administered and, for children who make 0 to 6 errors, an additional ceiling set of 10 items is administered. Simulations showed that approximately two-thirds (67%) of children would require only the core sets, 7% would require the core and basal sets, and 26% would require the core and ceiling sets; this distribution was considered in the selection of items.

4 Imputations to solve missing data problems in Wave 2.5

A number of the variables in the Wave 2.5 data files have higher levels of missingness than is usual for LSAC self-complete questionnaires. This chapter details the imputations made, and also those considered and rejected, in order to limit the amount of missing data. Using answers to other questions could impute some of these "missings", and this was done wherever possible.

Examination of the questionnaires revealed the following main reasons for the high levels of missing data (note that many of these are not exclusive to Wave 2.5 but appear to be exacerbated by other problems with the Wave 2.5 questionnaire):

1. Formatting issues. On pages where questions were in two columns at the top of the page but then in only one column at the bottom of the page, some respondents missed the second column at the top of the page. This affected the following questions:

Cohort	Question no.	Description
B	2	TV/computer in other rooms
B	3	Electronic games system
K	2(a)	TV/computer in other rooms
K	3(a)	Electronic games system, mobile, iPod
K	22	Required to look for work
K	23	Partner working
K	24	Income

Note: Number of missings was exacerbated by the instruction that appears under Q1 that said "If you do not have a computer at home, go to Q6".

2. Instructions to skip questions. There were a number of questions where the lead-in instructions requested that only people with particular characteristics (e.g. people who are currently working) complete the following question. It appears that some respondents may have skipped reading the preamble and made their own decisions about whether the question was relevant to them. Where this inconsistency led to people answering the question who should not have, their data was removed. However, missing data could not be replaced in most cases. This affected the following questions:

Cohort	Question no.	Description
B	23	Main reasons not in paid work
B	24	Plans about paid work
B	31	Effect of government benefits on attitudes to work
B	32	Attitudes to work
K	4	Computer use at home
K	5	Internet use at home
K	26	Effect of government benefits on attitudes to work
K	27	Effect of work on school involvement
K	28	Attitudes to work

3. "None of the above" questions. A number of questions provided a "none of the above" option if none of the other response categories applied. Experience with other self-complete forms has shown that it is not uncommon for people to omit ticking "none of the above". In general, it could be assumed that many of the responses to questions that had no response categories ticked are in fact "none of the above"; however, some people may skip questions for reasons that are not readily apparent. This affected the following questions:

Cohort	Question no.	Description
B	2	TV/computer in other rooms
B	11	Child care
B	18	Life events
B	24	Plans for paid work
B	25	Government benefits
B	28	Current study, etc.
B	35	Child support arrangement services
B	43	Help from other parent
K	1	TV/computer in child's bedroom
K	2	TV/computer in other rooms
K	3	Electronic games system, mobile, iPod
K	11	Child care
K	16	Life events
K	21	Government pensions
K	31	Child support arrangement services
K	39	Help from other parent

4. "Yes/No" questions. As with "'none of the above", it seems that some of the missing data for these questions could be explained by respondents for whom the "no" response was relevant omitting to tick the "no" option. In addition, if the question included a "go to" instruction, then sometimes respondents forgot to tick whichever option they selected. The main questions affected are:

Cohort	Question no.	Description
B	16	Is study child the youngest?
B	19	Do you currently have a paid job?
B	33	Does the study child have a PLE?
K	14	Is study child the youngest?
K	17	Do you currently have a paid job?
K	20	Are you currently looking for work?
K	22	Are you required to look for work?
K	29	Does the study child have a PLE?

5. Questions where '0' is a valid response. These are often left blank as respondents feel they don't apply. This affected Q14 for the B-cohort (number of changes to childcare arrangements) particularly since those currently without arrangements had been instructed to skip the previous two questions.
6. There were a number of cases in the B cohort (7) and the K cohort (29) that had roughly 90% of missing data. These cases have been excluded from the raw data files and the final files.

4.1 Rationale for imputations

4.1.1 Presence of media devices in the home and amount of time spent using these devices (B cohort Q2 to Q4)

An attempt was made to impute whether a child had access to the facilities listed in these questions by whether they had reported using them at Q4. Unlike many of the other imputations mentioned in this chapter, respondents were expected to answer the subsequent question regardless of their response to the previous ones. This meant some meaningful checks of the concordance between responses were possible.

However, most of the children (30 out of 41) who don't watch any TV still have TV in the home, so it couldn't be assumed that if children don't watch TV, they don't have access to one.

In Q1 and Q2 on the devices in their home, 46 of the 52 children were reported as watching some television in the home, even though these questions indicated they did not have a TV. While this indicates a misinterpretation of at least one of the questions, it would appear that the presence of a television in the home couldn't be reliably inferred from a response indicating that the child watches television in the home.

So therefore we can neither confirm nor refute the presence of a television in the home from the response to Q4a.

Likewise, 268 out of the 384 children that don't use a computer still indicated that they have one in the home, and 30 of the 51 children without a computer still use one. Again, the correspondence between the items isn't reliable enough to impute a response on whether there is a computer in the home.

For respondents with non-missing data for both the device ownership question (Q3) and the amount of time spent playing computer games (Q4), only in 25 of the 362 cases where the child doesn't have an electronic games system do they play one. Also, in 385 of the 435 cases where the child does have an electronic games system, they spend some time playing it. Since this data follows the basic correspondence that would be expected, the presence or absence of a console has been imputed by whether the child plays with one when this information is missing. This has added 141 "no" responses and 107 "yes" responses. It has also been imputed that if the child doesn't have access to an electronic games system at home, the time spent playing with one at home will be nil, altering 1,541 responses from missing.

4.1.2 Devices in the home and possession of personal devices (K cohort Q2 and Q3)

No checks are possible on television use, electronic games systems or iPods. For mobile phones, there is no implication in the "use of mobile phones" items at Q6 that the child has to use their own mobile phone to do these things, so ownership of a mobile phone can't be imputed.

For computers and the Internet, cases where the child has a computer in their room are actually a little less likely to have one somewhere else in the home (87% vs 93%), so it can't be assumed that if they have one they'll have the other. Generally, if the respondent has given good answers to Q4 and Q5, they do have a computer in the home, but it's not necessarily the case that because the respondent has answered these questions incompletely that they don't have a computer. Therefore, it's impossible to impute accurately.

4.1.3 Presence or absence of child care (B cohort Q11, K cohort Q11)

If the respondent indicated that the child did spend time at child care, the child was assumed to have an "other" type of child care for the K cohort; however, for the B cohort, they were imputed as having child care but the type of child care was set to missing since no "other" option was available. This affected three cases in the B cohort and two cases in the K cohort. If the respondent reported zero for the number of days or hours per week of child care, it has been imputed that the child had no child care. This affected one case for the B cohort and two cases for the K cohort.

4.1.4 Is the study child the youngest child in the home (B cohort Q16, K cohort Q14)

The study child was aged either 3-4 years for the B cohort or 7-8 years for the K cohort at the time of the Wave 2.5 questionnaire. Therefore, if the respondent indicated that the age of their youngest child corresponds with this, it has been imputed that this is the study child, affecting six cases in the B cohort and 14 cases in the K cohort. If the age given was younger than this, it has been imputed that this wasn't the study child, affecting 14 cases in the B cohort and four cases in the K cohort.

4.1.5 Does the respondent have a paid job (B cohort Q19, K cohort Q17)

If the respondent indicated that they did work some hours then they were imputed to have a paid job, affecting two cases for the B cohort and 13 for the K cohort that were previously missing. If they said they generally work zero hours then they were imputed to have no job, affecting one case for the B cohort. If B cohort respondents were missing data for work hours and their desired number of work hours, but had a

response for why they were not currently working or for their future work plans, they were imputed as being out of work, affecting four cases for the B cohort. This question was not asked for the K cohort so no similar imputation was possible.

Some of the remaining missing cases had data for the desired work hours question; however, those with or without a job could logically answer this question, so this provided little indication of the true response to whether they were working. The attitude items for those in work (Q32 for the B cohort, Q27 and Q28 for the K cohort) could also be answered by some non-workers on the basis of previous work experience, so imputation based on responses to these was deemed unreliable.

4.1.6 Whether government benefits are received (B cohort Q25, K cohort Q21)

For the K cohort, if the respondent indicated that they are required to do an activity test, it could be imputed that a benefit is received; however, none of the missing cases met this criteria. The only other possibility for imputing this question would be to look at the effect of government benefits on work-plan items (B cohort Q31, K cohort Q26); however, there is no way of knowing if the respondents that didn't answer these questions were getting family tax benefits. Also, the skip is not very well highlighted in the formatting, so there can be little confidence that those who answered the question understood who it was for.

4.1.7 Whether the study child has a PLE (B cohort Q33, K cohort Q29)

A number of the missing cases have been classified on a case-by-case basis based on responses to the follow-up questions on child support. The criteria for these classifications involved the amount of missing data, the amount of data that might indicate the presence of a PLE (e.g. having a child support arrangement vs not having one), as well as whether a PLE was present at Wave 2. This created one extra "yes" response and one extra "no" response for the B cohort and four extra "no" cases for the K cohort. After this process there were 16 cases that were missing all subsequent information for the B cohort and nine for the K cohort. Most children do not have a PLE; however, some of these cases could be from people who gave up on the questionnaire. It can be reasonably assumed that if they answered 50% of items in the most recent question required of them then they haven't given up on the survey and therefore are just cases without a PLE. This added an extra 12 "no" cases for the B cohort and three for the K cohort.

4.1.8 Respondent information (B cohort Q47)

Initially, there were 100 records that were missing the respondent information ("who completed this form?") in the B-cohort file and 85 records in the K-cohort file. ABS were able to correct 88 B-cohort records and 69 K-cohort records by matching the names of the people that completed the Wave 2.5 form to the names of people who participated in Wave 2. The location of this question may have been a factor in why there are missings, because the question is located at the bottom of the back of the form.

5 Review of main educational program of 4-5 year olds

5.1 K cohort

In investigating the quality of the data for the child's educational program type (*cpc06a4)* at Wave 1, concerns were raised in regard to the consistency of responses to this item with other information from the face-to-face interview and the teacher questionnaire. It was decided to provide a corrected version of *cpc06a4* as well as the original version. The correction involved two processes:

1. If teacher data was present and contradicted the value given by the Parent 1, the value indicated by the teacher data was used instead; or
2. If no teacher data was present, a number of checks were performed on the consistency of the parent's response with other data given (e.g. number of hours in care, the age of the study child, etc.). If a majority of cases with teacher data were corrected when they had the same combination of the original response and number of inconsistencies, then those without teacher data were corrected to the majority value. For example, it was found that among those cases with two or more inconsistencies whose original response was "Pre-year 1 in a school", more than 50% of the teacher data, where available, indicated that the true response was "Preschool in a school". This value was therefore assumed to be most likely for these cases in the absence of teacher data.

More information on this process is provided in the Data User Guide (available from www.growingupinaustralia.gov.au/data/docs/userguide/index.html).

At Wave 3, respondents were asked to confirm the details of the educational program the child was in at the time of the Wave 2 interview two years prior, and were then asked about the details of the child's educational programs from ages 3 to up to 6 years prior (working backwards until either the child wasn't in an educational program or was in pre-school/kindergarten).

Table 11shows the information captured for each year. This section suggests improvements to the imputation based on this new data.

To determine how to best use this data, some determination has to be made as to its quality. As an initial check, the recall data was checked for reliability with itself. The data was considered unreliable if there was a greater gap in year level than the number of years between time points, or a lesser gap unless there was an indication that a year level was repeated. This check revealed 14.5% of the cases were unreliable. The data collected at Wave 3 for these cases was not used to impute *cpc06a4*.

The data was then examined to quantify the number of inconsistencies with other data items from the Wave 1 questionnaire. The following circumstances were considered to be inconsistent:

- The child was in a "pre-year 1 program" at school and was:
  - attending this program fewer than 5 days/week; or
  - attending this program less than 30 hours/week; or
  - younger than 55 months of age at Wave 1; or
  - in "Year 1" in Wave 2 unless indicated they had repeated a grade level.
- The child was attending a "pre-school" (other than in day care) and was:
  - attending this program for 30+ hours/week; or
  - more than 62 months of age at Wave 1; or
  - in "Year 2" at Wave 2.

Table 11: Variables capturing previous years educational programs for the K cohort at Wave 3	
Questions	
1) What program did child attend the year before, that is in (3 years prior)?	
Year 1 (Grade 1)	→ 3
Pre-year 1 program	→ 3
Preschool/kindergarten program	→ 3
Long day care	→ 3
Home-schooled	→ 3
Other	→ 2
Child did not attend an educational program	→ End of recall items
2) Other specify	→ Previous year
3) Was that located in a school?	
Yes	→Previous year
No	→ epc59d?
4) Was it a...?	
Preschool/kindergarten only centre	→ End of recall items
Preschool/kindergarten in a long day care centre	→ End of recall items
Mobile pre-school	→ End of recall items
Long day care centre	→ End of recall items
Other	→ End of recall items

Three different versions of this information were compared using these checks:

1. "Original"—the original value entered from the Wave 1 face-to-face interview.
2. "Teacher"—the original data corrected when it disagreed with data obtained from the Wave 1 teacher questionnaire.
3. "Recall"—the information as recalled by the respondents at Wave 3 for 4 years prior.

Among those cases that had teacher data at Wave 1 and had reliable recall data at Wave 3,[2] 81% were found to have no inconsistencies when using the recall data. This compares with 65% with no inconsistencies using the original data and 84% when using the teacher data. So it would seem that the teacher data is still the most consistent indicator of the true value: however, the recall data is also reasonably consistent.

In order to determine how to best use this data in the imputation, two different methods were tried. In the first, the recall data was substituted for the original data automatically. There was agreement between the value created using this scheme and the one using the teacher questionnaire data in 76% of cases.

For the second approach to imputation, the recall data (when reliable) were used as an additional check to those listed above and imputations were made on the basis of the number of unlikely combinations of data.

Under the second scheme, the following corrections were made:

1. Children in "Year 1" were automatically recoded to "pre-year 1".
2. Children in "pre-year 1" with two or more inconsistencies were recoded to "preschool in a school".
3. Children attending a "preschool in a school" with two or more inconsistencies were recoded to "pre-year 1".
4. Children attending a "preschool at a non-school centre" with two or more inconsistencies were recoded as being in a "day care centre with a preschool program".

[2] That is, minus the 14.5%mentioned above.

For cases with Wave 1 teacher data, the data generated by these corrections matched the teacher data in 80% of cases, better than using the recall data by itself, and better than the correction scheme used prior to the recall data becoming available (which matched in 73% of cases). This approach has therefore been taken.

5.2 B cohort

Given the problems experienced for the K cohort at Wave 1, a different set of questions on educational programs was developed for the B cohort at Wave 3 (see Table 12). In Wave 1 for the K cohort, the data collected from the face-to-face interview on educational programs differed from that collected in the teacher questionnaire in 29% of cases. In Wave 3, for the B cohort, there were differences in 13% of cases.

However, when the consistency of the teacher data and the parent data was tested against other answers in the Wave 3 interview, it was found that neither version had many inconsistencies; however, the teacher corrected version had slightly more (3% versus 2.7%).

In the seven cases (so far) with inconsistencies when the teacher data was used, the teacher's response was "pre-year 1 school program" while the parent's was "preschool program in a school". These cases may represent programs that don't fall neatly into either category (e.g. classes at a pre-year 1 level that children attend part-time), although there is no consistency in terms of state of residence of the children or the organisational basis of the school (e.g. independent versus state versus Catholic). Whatever the situation is with these cases, there seems to be little reason to correct the parent data or teacher data when there is little indication of which is correct.

Outcomes:

1. *Teacher data still to be used to correct parent data when available in determining educational program at Wave 1 for the K cohort.*
2. *Recall data to be used as an extra consistency check within the existing process when imputing this information when teacher data is absent.*
3. *No imputation to be performed on Wave 3 B-cohort educational program data.*

Table 12: Variables capturing current educational programs for the B cohort at Wave 3

Questions	
1) (Thinking about the arrangement the child uses for the most hours per week) is this located in a school?	
Yes	→ 1
No	→ 5
2) What class or program does child attend?	
Year 1 (Grade 1)	→ 4
Pre-year 1 program	→ 4
Preschool/kindergarten program	→ 4
Long day care	→ 4
Other, e.g. multi age classes, early intervention	→ 3
3) Other specify	→ 4
4) Does child attend this program at	
A government school?	→ Further items
A Catholic school?	→ Further items
An independent or private school?	→ Further items
5) Which of the following best describes where child goes?	
Preschool/kindergarten only centre	→ Further items
Preschool/kindergarten in a long day care centre	→ Further items
Mobile preschool	→ Further items
Long day care centre	→ Further items
Other	→ 6
6) Other specify	→ Further items

6 Cleaning of income data

Following the original release of the data, users reported problems with outlying values in the continuous income variables (i.e. afn09a, afn09b, afn09m, afn09f, cfn09a, cfn09b, cfn09m, cfn09f). While this is not unusual for income, it appeared that some of these cases had unusual responses to other questions for those with such high incomes (e.g. more modest incomes reported when asked about combined yearly income at K20 of the face-to-face interview, more menial occupations). It appears that many of these are due to discrepancies between amount and time period when reporting income (e.g. giving yearly income as weekly). Many of these outliers have been subsequently cleaned up, although certain assumptions have been made to do so.

The process for cleaning the Wave 1 data used adaptations of the data query rules coded into the Wave 2 CAPI instrument. As well as providing a logical framework to underpin the investigation, this will also help in making the data more consistent longitudinally.

The rules used were as follows:

1. If a respondent's only source of income is government benefits or salary they should not report an income of $0 or a loss.
2. If profit or loss is a source of income then incomes >$200,000/year should be queried unless they also have a salary.
3. Where government benefits are the main source of income, incomes >$750 a week should be queried.
4. For all other combinations of income types, incomes >$260,000/year should be queried.

Cases identified by the first of these rules were all set to missing. Most of these seem to be due to respondents not counting government benefits as income. In the B-cohort file there were 49 cases of this for Parent 1 and 6 for Parent 2, while for the K cohort, 31 cases were identified for Parent 1 and 6 cases were identified for Parent 2.

For those identified by the other three rules, if the categorical annual income for Parent 1 and Parent 2 at K20 was consistent with the continuous values they were left as is. If there was an obvious correction that could be applied (e.g. deleting a zero from an income figure, changing the time period from weeks to year) to bring the income into or close to the range specified at K20 then this was applied. If there was no way that the continuous income values could be made reasonably consistent with the combined parental yearly income then the response was assumed to be an error and was made missing.

Restrictions on publishing case level information limit what can be disclosed about these cases. However, for Parent 1s in the B cohort, of the 31 cases identified by rules 2 to 4, six were made missing, eight were corrected and 17 were left as is. For B-cohort Parent 2s, of the 48 cases identified, four were made missing, 10 were corrected and 34 were left as is. For K-cohort Parent 1s, of the 30 cases identified, five were made missing, six were corrected and 19 cases were left as is. For K-cohort Parent 2s, of the 42 cases identified, five were made missing, eight were corrected and 29 were left as is.

In Waves 2 and 3, suspicious cases were identified using the above rules. These cases were checked against their income data from earlier waves, plus other information such as work hours and occupation. As would be expected, data collected with the CAPI instrument was cleaner, and fewer imputations had to be made. In Wave 2, seven corrections were made to Parent 1 income, four corrections to Parent 2 income, and one to the income of other adults in the home. In Wave 3, only one Parent 1 and one Parent 2 required correction.

7 Height differences

In the leave-behind questionnaires for both parents at Wave 1 and Wave 2, the parents were asked to report their height and weight so their body mass index (BMI) could be calculated. In cleaning Wave 2 data, it was discovered that there was a large number of discrepancies between the values reported by the same people at Wave 1 and at Wave 2. In fact, only 50% of respondents reported a value that was within 1% of their Wave 1 value.

Further investigation failed to find any explanation other than respondent error for the vast majority of these cases. In order that data analysts could assume that any observed changes in BMI were due to changes in reports of weight rather than height, it was decided to impute the value of height to be the average of the two reported values.

At Wave 3, the question on the Parent 1's height was asked of all new Parent 1's and those that had not returned a self-complete form at Wave 2, plus a handful of cases where Parent 1 had swapped places with Parent 2. However, for Parent 2 the height data was still collected by self-complete form, so sequencing cases around the question was not an option. Hence, for many,[3] there are now three points of data collection.

When the study child's height is measured as part of the interview process, a third measurement is taken if the first two disagree by more than 0.5 cm. If this is the case, the estimate of the child's height is considered to be the average of the two that correspond the most closely. This method of estimation means that the least reliable estimate has no effect on the result. It is suggested that in cases with three data points for a parent's height, the "clean" result provided on the data file could similarly be the average of the closest two responses. As is done currently, the values of parental height for each wave prior to this cleaning will remain on the data file if analysts wish to use their own approach.

Figure 7 shows the discrepancy between the two values used to create the "clean" result for those parents with two data points versus those with three. Those with three data points had two that agreed in 77% of cases. Those with two data points had agreement in only 42% of cases. It should be noted, however, that at Wave 2, 45% of cases had agreement between the two data points, so there is some evidence that those who were more likely to return self-complete questionnaires were more likely to give accurate data.

It was decided on inspection of Figure 7 that any case with more than a 10 cm discrepancy between the two closest values should be considered unreliable and therefore should be set to missing. This would affect 4% of Wave 3 parents with two data points and less than 0.1% of cases with three data points.

Outcome:

1. *This problem with the height data was presented to the February 2009 data Expert Reference Group Meeting and the group decided that if the differences are less than 10 cm then average all three, otherwise average the closest two. Consequently, this is how the height data has been adjusted.*

[3] 72% of those who returned Parent 2 questionnaires at Wave 3 have data from all three waves. Of all Wave 3 Parent 2s, 11% had no height data, 16% had one data point, 22% had two data points and 51% had 3 data points.

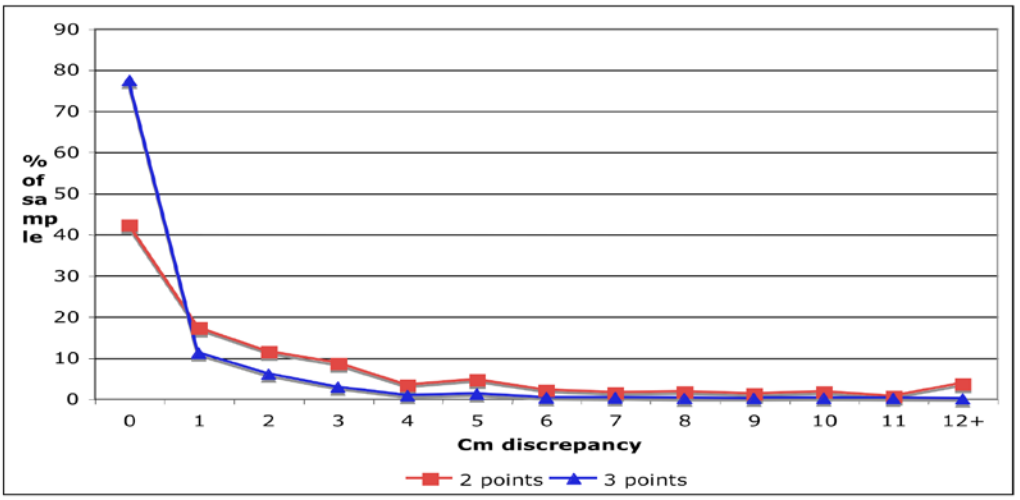

Figure 7: Centimetre discrepancy in two closest data points for those with three vs two data points on parental height for Wave 3 respondents

8 Data issues in Wave 3.5

8.1 Q30B and Q24K

These questions ask how many hours per week the study child spends doing each of the following activities: watching TV, watching DVDs, using the computer, playing games, and listening to music. A total of 20 B and K respondents indicated an unusually large number of hours for some of these activities (that were not able to be corrected through normal editing processes). If the respondent indicated that the child spent more than 40 hours on weekdays doing an activity, these activities were set to missing.

This affected:

- nine cases for the B cohort and two cases for the K cohort for time spent watching TV on weekdays
- one K-cohort case for time spent listening to music.

If the respondent indicated that the child spent more than 20 hours on a weekend doing an activity, these activities were also set to missing. This affected:

- six cases for both the B cohort and the K cohort in relation to time spent watching TV on the weekend
- two B-cohort cases and three K-cohort cases for time spent watching DVDs on the weekend
- two B-cohort and eight K-cohort cases for time spent using a computer on the weekend
- one B-cohort and three K-cohort cases for time spent playing a game console on the weekend
- one case for both the B cohort and the K cohort in relation to time spent listening to music on the weekend.

Care should be taken when using data on media use as there is no provision in the form for parents to report whether the activities were undertaken concurrently. For example, the TV may be on and the child may also be using the computer; therefore, it may be acceptable that some reports of total activities are greater than the total hours in a weekend. However, while watching TV and using a computer at the same time may be plausible, it cannot be clearly determined that this is what is occurring from the responses given.

8.2 Q13B and Q35K

These questions ask the number of days per week during school term that the study child walks, rides, uses public transport or goes by car to and from school. Although respondents were asked to report the MAIN form of transport each day, some respondents reported multiple travel types.

The question aimed to obtain the main mode of transport used during an average school week to and from school. Therefore, the number of trips to and from school should sum to five each. However, this was not the case for some respondents, as shown below.

B cohort	To school	From school
Less than 5	125	132
5	2,694	2,785
More than 5	126	116
Total	2,945	2,901

K cohort	To school	From school
Less than 5	35	77
5	2,739	2,674
More than 5	183	181
Total	2,957	2,932

The cases in which the number of trips to or from school did not add to five have been set to missing. This affected 296 in the B cohort and 319 in the K cohort in total. Two hundred and three B-cohort and 157 K-

cohort cases had this problem for both to and from school. Forty-eight B-cohort and 61 K-cohort cases had a problem with to school, while 45 in the B cohort and 101 in the K cohort had a problem with from school.

8.3 Q25B/Q26B and Q12K/Q13K

8.3.1 Instructions to skip questions

There were a number of questions where respondents did not follow the skip instructions correctly. Where this inconsistency led to people answering the question who should not have, their data was removed. However, missing data could not be replaced in most cases. This affected the following questions:

Cohort	Question no.	Description
B	5	Looked forward to going to school
B	6	Upset or reluctant to go to school
B	8	Teacher informs parent of child's progress behaviour
B	9	How well child gets along with other children
B	10	Quality of education
B	11	Teacher/Parent relationship
B	12	School social capital
B	13	Travel to and from school
B	14	Distance to school
B	15	Provide help with homework
B	16	Experiences before child started school
B	17	Communication with school before child started
B	18	Child's transition to school
K	20	Child began to menstruate

8.3.2 "None of the above" questions

A number of questions provided a "none of the above" option if none of the other response categories applied. Experience with other self-complete forms shows that it is not uncommon for people to omit ticking "none of the above". In general, it could be assumed that many of the responses to questions that had no response categories ticked are in fact "none of the above". However, some people may skip questions for reasons that are not readily apparent. In these cases, the item has been left as missing. This affected the following questions:

Cohort	Question no.	Description
B	16	Experiences before child started school
B	17	Communication with school before child started
B	27	Sleep problems
B	28	Stressful life events
K	21	Stressful life events

8.3.3 "Yes/ No" questions

As for "none of the above", it seems that some of the missing data for these questions could be explained by respondents for whom the "no" response was relevant, omitting to tick the "no" option. The main questions affected are:

Cohort	Question no.	Description
B	33	Family rules
K	6	Family rules about homework
K	7	Special place to do homework
K	28	Family rules

8.3.4 Substantial amounts of missing data

There were some cases in the B cohort (1) and the K cohort (5) that had roughly 90% or more missing data. These cases have been excluded from the raw data files and the final files.

9 Data issues in Wave 4

9.1 Instances of child but not parent participation

Ideally, it is expected that parents who provide consent to interview their children would agree to be interviewed themselves. However, this does not happen in 100% of cases. In Wave 3, there was only one case out of 8,718 home visits in which a parent was not interviewed but provided consent to interview a child. In Wave 4, there were five out of 8,405 cases where parents were not interviewed but agreed to their children being interviewed. The main reasons for parents to refuse a home interview but allow their child to be interviewed were lack of interest and time.

Parent 1 refusal to the home interview might lead to missing household information.[4] If household information is not available this record is not included in the household file. As a result, in Wave 4, the household file hhgrk10 contains 4,164 records and main file lsacgrk10 contains 4,169 records. If data users intend to merge these data sets they need to be aware that there is a mismatch between the datasets for five cases.

To help a data user identify cases with available household data the following variables were created *hhresp for both cohorts. In variable names the asterisk refers to an age indicator (In Wave 4, d refers to B cohort and f refers to K cohort).

While in Waves 3 and 4 the discrepancy in child and parent participation is minor, it might increase in future waves due to: (1) changes to the interview procedure; or (2) children becoming more active participants/refusers in the study; or (3) increases in other activities of parents and children meaning fewer times when both are at home at the same time and available for interview.

9.1.1 Changes to the interview procedure

From Wave 4, interviewers were provided with two laptops and were able to conduct “parallel interviewing”. The interview was split into two streams; all the Parent 1 (P1) questions were on laptop 1 and all the study child (SC) questions were on laptop 2. As a result, the interviewer had the flexibility to complete child and parent interviews either at the same time or at different times.

9.1.2 Children becoming active participants/refusers of the study

All of the cases where children were interviewed but parents were not belong to the K cohort in both Waves 3 and 4.

9.2 ACASI

The audio computer-assisted self-interview (ACASI) contains questions that are skipped if the study child has no mother and/or father figure in their life or does not attend school. As these circumstances were determined in the CATI component of the Parent 1 interview on laptop 1, they were not apparent in the interview on laptop 2 when the ACASI was conducted. Consequently, in order for the relevant questions to be skipped, prior to providing the laptop 2 to a study child to complete the ACASI, the interviewer was required to enter into the ACASI instrument whether there was “Mum” and/or “Dad” figures in the study child’s life and whether the study child had been attending school. The interviewers were asked to use their knowledge of the family that they gained after completing the CATI component with the Parent 1.

When deciding if it would be appropriate to ask about a “Mum” or “Dad” the interviewers were asked to be sensitive towards the situation of the child as family structure can be complicated. The interviewers were instructed that if they were unsure whether there was a “Mum” or “Dad” in the child's life and/or whether a

[4] If a Parent 1 completes CATI prior to the home interview, household information is not missing.

child had any contact with them to enter “no father” or “no mother” so as to not distress the child. For example, in the situation where the current Parent 1 and Parent 2 were not the biological parents, it was unclear as to who the child would be referring when asked about their mother or father. In just a few cases there was an interviewer’s error and wrong information was rolled into the ACASI module.

In just a few cases there were inconsistencies between the household information and the interviewer’s assessment of whether there was a mother in the child’s life. As a result, there were 10 cases for which a mother was recorded in the household file but where all questions about the mother were skipped in the ACASI module. There were also 32 cases where all questions about the father in the ACASI module were skipped but a father was identified in the study child’s life in the household file. To identify these problematic cases the following variables were created (asterisk refers to an age indicator):

- *mumsk—a mother is identified in the household but questions about “Mum” figure are skipped in ACASI module
- *dadsk—a father is identified in the household but questions about “Dad” figure are skipped in ACASI module
- *schsk—child is in school but questions about school are skipped in ACASI module.

With regards to the school attendance, 40 children who were in school (as identified in the education module) skipped all the questions about school in the ACASI module. This mismatch was mainly due to manual errors and one of the problems of the method used, that is, information about the child’s education is provided by Parent 1 and stored on the laptop 1 and the study child completes ACASI separately on the laptop 2. This information being stored on different computers means that these instruments do not talk to each other.

9.3 Matrix Reasoning

Matrix Reasoning (MR) is a test from the Wechsler Intelligence Scale for Children, 4th edition (WISC-IV) (Wechsler, 2004) for ages 6-7, 8-9 and 10-11 years. This test of non-verbal intelligence presents a child with an incomplete set (later referred to as item) of pictures and requires the child to select the picture that completes the set from five different options. The instrument is comprised of 35 items of increasing difficulty.

9.3.1 Administration rules

According to the WISC-IV manual, the administration of Matrix Reasoning should follow a set of rules. We are not going to discuss all the rules in detail but rather focus on the rules crucial for our purposes.

Administration of the test should start at the age-specific start-point, which is indicated in the WISC-IV manual. Item 4 is the start-point for children aged 6-7 (B cohort) and Item 7 is the start-point for children aged 10-11 (K cohort).

Items prior to age-appropriate start-points are called reversal items. Reversal items are asked only if a child provides incorrect answers on the first or second start-point item. If a child answers incorrectly either of the first two items from the start point, the interviewer asks the preceding items (reversal items) in reverse sequence until the child answers correctly two consecutive items and then goes back to the age-appropriate items and proceeds with the rest of the test. This is called reverse administration. For example, if a 6-year-old child answered correctly on Item 4 and incorrectly on Item 5, an interviewer should reverse to Item 3, then Item 2. If Item 3 or Item 2 is incorrect then Item 1 is administered. If Items 3 and 2 are correct, Item 1 is not administered. After administering reversal items, the interviewer goes back to Item 6 and proceeds with the rest of the test.

9.3.2 Scoring rules

The total raw score of MR is equal to the number of correct items starting from an age-appropriate start-point plus the total score on the reversal items. For items administered from the age-appropriate start-point a raw score of 1 is assigned for each correct answer.

For reversal items the following scoring rules are applied:

Rule 1—Each reversal item gets a score of 1 if the reverse administration is not required (first two items from the start-point are answered correctly). For example, if a 6-year-old child answers correctly Items 4 and 5, the reversal Items 1, 2 and 3 are scored 1 each.

Rule 2—Each reversal item gets a score of 1 if a child correctly answers two consecutive reversal items. For example, if a 6-year-old child answers correctly on reversal Items 3 and 2 and Item 1 is not administered or answers incorrectly on Item 3 and correctly on Items 2 and 1, the reversal Items 1, 2 and 3 are scored 1 each.

Rule 3—Each correctly answered reversal item gets a score of 1 and each incorrectly answered reversal item gets a score of 0 if a child does not answer correctly on any two consecutive reversal items. For example, a 6-year-old child answers incorrectly on the reversal Items 3 and 1 but correctly on the reversal Item 2. Then, Items 3 and 1 are scored as 1 each and Item 2 is scored as 0.

9.3.3 Administration of MR in LSAC

Due to the technical difficulties in programming, the reverse administration was not implemented in the LSAC MR instrument, that is, if LSAC children answered either of the two items from the start-point incorrectly the reversal items were never administered. Table 13 shows a number of cases where first two items from an age appropriate start-point were answered correctly and incorrectly for B and K cohorts.

Table 13: Frequencies of correct responses on the start-point items

	N	%
B cohort		
Item 4 and Item 5 are correct	3,964	95
Item 4 or Item 5 is incorrect	216	5
Total	4,180	100
K cohort		
Item 7 and Item 8 are correct	3,908	95
Item 7 or Item 8 is incorrect	195	5
Total	4,103	100

It can be seen from Table 1 that 95% of children answered the first two items from the age-appropriate start-point correctly and did not require the reverse administration. The raw scoring for these children was based on Rule 1. The other 5% of children answered one of the first two administered items incorrectly and, therefore, required the reverse administration to identify which rule for scoring should be used, Rule 2 or Rule 3. Given that the reverse administration was not available, it was decided to assign all reversal items a raw score of 1 regardless of whether the first two administered items were answered correctly or not. As a consequence, some of the 5% of children might have had their MR scores overestimated. The following variable was created to identify these 216 cases in the B cohort and 195 cases in the K cohort:

$$*mrrawi = \begin{cases} 1, \text{if either of one first two items from start point is incorrect} \\ 0, \text{otherwise} \end{cases},$$

where * refers to appropriate age indicator.

The MR scores on Items 1 to 6 from previous waves are examined below.

K cohort

Out of 195 children from the K cohort who did not answer either one of two first items from the start-point at Wave 4, 185 children did MR at Wave 3. While the reverse administration was not implemented in Wave 3, all items were administered, that is, Item 1 was the first administered item. This allows us to cross-check how many children out of 185 gave two consecutive correct answers on Items 1, 2, 3, 4, 5 and 6. There were 179 who answered correctly either on all Items 1, 2, 3, 4, 5 and 6 or answered correctly on two consecutive items. In this instance, at Wave 3, they were assigned the maximum possible score. Assuming that cognitive ability of children remains stable over time, we would expect these children would obtain the maximum possible score for the first 6 items at Wave 4 too.

B cohort

In Wave 4, B-cohort children were administered the MR test for the first time. However, in Wave 2, K-cohort children did the MR test and they were the same age as the B-cohort children in Wave 4. Therefore, the relative comparison could be made against the K-cohort children of the same age. In Wave 2, there were 269 (6%) of K-cohort children who answered Items 4 or/and 5 incorrectly. Out of these 269 children, only 16 children did not answer correctly Items 3 and 2 or Items 2 and 1.

Therefore, based on the data from previous waves, we would expect only a very small number of children in either cohort to have their MR ability overestimated through the changes in administration and scoring.

10 Data issues in Wave 5

10.1 Geography

The first four waves of LSAC data included geography items such as postcodes and various levels of the Australian Standard Geographical Classification (ASGC) that were generated from geocoding of the residential addresses of study families. In Wave 1, the geocodes were based on global positioning system (GPS) coordinates obtained by I-view interviewers at the time of interview, while in Waves 2 to 4, they were based on residential addresses collected by Australian Bureau of Statistics (ABS) interviewers.

In July 2011, the ABS introduced a new statistical geography framework called the Australian Statistical Geography Standard (ASGS) to replace the ASGC. The main purpose of the ASGS is to disseminate geographically classified statistics. It provides a common framework of statistical geography enabling the publication of statistics that are comparable and spatially integrated.

Improved data sources and technology have allowed the ABS the opportunity to create a better geography optimised for the release of ABS statistics. A new robust and stable structure means that changes over time are minimised, assisting in the maintenance of quality time-series data. In addition, the ASGS, together with improved methods of calculation, allows for more accurate correspondences to translate ABS data to non-ABS administrative and geographic regions.

For further information on this new standard refer to 1270.0.55.001—Australian Statistical Geography Standard (ASGS): Volume 1—Main Structure and Greater Capital City Statistical Areas, July 2011.

To take advantage of this more comprehensive, flexible and consistent way of defining Australia's statistical geography, the ASGS will be included from Wave 5 onwards. To ensure that there is a common geographical standard across waves, the decision was made to:

- dual-code Wave 5 residential addresses to ASGC and ASGS, enabling comparison of old and new classifications; and
- back-code Waves 1–4 residential addresses to the new standard ASGS.

The new variables added to the general release file for each wave are shown in Table 14.

Table 14: New geography variables included from Wave 5

Without age variable name	Label
gccsa	Australian Statistical Geography Standard (ASGS)—Edition 2011—Greater Capital City Statistical Area Structure
sos	Australian Statistical Geography Standard (ASGS)—Edition 2011—Section of State
sa22011	Australian Statistical Geography Standard (ASGS)—Edition 2011—SA2
sa32011	Australian Statistical Geography Standard (ASGS)—Edition 2011—SA3
sa42011	Australian Statistical Geography Standard (ASGS)—Edition 2011—SA4
absra	Australian Statistical Geography Standard (ASGS)—Edition 2011—Remoteness Area (ABS)

Most addresses were auto-coded using ASGS address coders, which allow addresses to be linked to geographical areas. However, in some cases, addresses were either incomplete, had spelling errors or, more rarely, were identical addresses in the same suburb. In these cases, addresses were manually cleaned to reduce the number of records with missing geocodes. After these steps, there were still some records unable to be geocoded to ASGS (level SA2). These numbers for Waves 1–5 are provided in Table 15.

Table 15: Number of records missing SA2 by wave

Wave	Number of responding records not coded to SA2
5	2
4	34
3	13
2	12
1	20

To enable coding to the ASGS, many addresses needed cleaning to ensure accurate data. As a result, some records have SLAs where there were none previously, and others have been coded to a different SLA.

The 2011 Census and SEIFA data are available in the new ASGS classifications. However, while it is possible to provide ASGS classifications for Waves 1 to 5, census and SEIFA data for 2001 and 2006 are not available for these new geographic classifications (ASGS).

From Wave 7 onwards only ASGS geography variables will be output on the files.

10.2 Occupation

LSAC data include variables for the occupation of Parent 1 (P1) and Parent 2 (P2). In recent waves, the occupation of Parent Living Elsewhere (PLE) and the parents of P1/P2/PLE (i.e. the study child's grandparents) are also included. These were coded using the Australian Standard Classification of Occupations (ASCO). The ANU4 scale—a scale of occupational status calculated using ASCO, which is an occupational classification system that classifies jobs according to skill level and skill specialisation—is also provided to data users for Waves 1 to 4.

Since Wave 2, LSAC occupation data has also been coded to the newer occupation standard, which is the Australian and New Zealand Standard Classification of Occupations (ANZSCO). ANZSCO was introduced in 2006 and was a product of a development program between the ABS, Statistics New Zealand and the Australian Government Department of Employment and Workplace Relations.

For further information on this standard, refer to 1220.0—ANZSCO—Australian and New Zealand Standard Classification of Occupations, First Edition, 2006.

The latest release of ASCO was in 1997, reducing its applicability to the current Australian workforce. Therefore, from Wave 5 onwards only, ANZSCO codes will be produced. To enable the transition to using ANZSCO, the study has:

- added ANZSCO codes to the Waves 2–4 data files, as these codes were already generated during these waves, and is investigating the possibility of providing ANZSCO for Wave 1 through correction code.
- replaced the ANU4 scale from Wave 5 onwards with the Australian Socioeconomic Index 2006 (AUSEI06) (McMillan, Beavis, & Jones, 2009), the latest in the series of occupation status scales developed by the ANU.
- provided AUSEI06 for Waves 2–4 and is investigating the possibility of adding to Wave 1 through correction code.

The new variables added to the general release file are in Table 16.

Table 16: New occupation variables included from Wave 5

Question ID	Label
pw08_5	Current occupation (ANZSCO code)
pw08_6	Current or most recent occupation (ANZSCO code)
pw08_7	Current occupation (AUSIE06 code)

The SEP variable (Z score for socio-economic position among all LSAC families) has been calculated from Waves 1 to 4 using ASCO classifications. Due to ASCO being unavailable for Wave 5, the SEP variable has not been calculated and hence is not available in the Wave 5 dataset. Further work will be done into ways we can calculate the SEP using the ANZSCO classifications and a new/revised SEP variable may be available in the future.

10.3 ACIR data issue (all waves)

After analysis of the ACIR data previously supplied, it came to light that immunisation rates in LSAC did not reflect national rates. After investigation with the data provider, it was found that data extraction up to Wave 5 had not extracted all the required records. This data has been rectified, however data users should not use the previous version of the ACIR data.

10.4 Changes to household files

10.4.1 Addition of "Person Type" to the files

In Wave 5, Person Type (f21a) is available on the Waves 1 to 5 files for the first time, with a code attached to each household member and wave. This item is derived from information collected in the P1 interview and amended where needed during processing. A list of the person types and a description of each is shown in Table 17.

Table 17: Person Type descriptors

Code	Person Type	Description
1	Study child	The study children are the focus of the study, and consist of two cohorts (B cohort aged 8–9 years and K cohort aged 12–13 years in Wave 5).
2	Parent 1	Parent or guardian who provides the greatest role in caring for the study child and is therefore likely to be the most reliable informant on the health, development and care of the study child. Parent 1 must live with the study child.
3	Parent 2	Study child's other resident parent/guardian, or the married or de facto partner of Parent 1. Another person in the household can be considered as Parent 2 if they are acting as a significant parental figure who helps to care for the child, and is a stable member of the child's residential family unit.
4	Usual resident	A person other than the study child and the study child's resident parent(s) who usually lives in the study child's house (e.g. siblings of the study child).
5	Non-resident	A person other than a parent who has previously been a resident of the household, but no longer lives in the same household as the study child.
6	Parent living elsewhere	A parent of the study child who does not live in the same household as Parent 1 and the study child. This person may previously have been a Parent 2 (or a Parent 1).
7	Temporary member	Includes people who, in-between waves, joined the study child's household for more than 3 months but have since left.
8	Empty row	In the household files row/member number 3 is always used for Parent 2 at Wave 1. When there was no P2 in the house at Wave 1, this row is left as an empty row. Also used when duplicate members are picked up.
9	Deceased	A person who was previously recorded as a resident of the household, but has died.

10.4.2 Changes in relationship to study child information for household members

For Waves 1 to 4, the household file carried forward the relationship to study child for each member in the household from Wave 1 or the subsequent wave for members entering the household after Wave 1. This means that for an existing household member, the relationship information in the household file is generally the same across waves. In some cases this will not reflect changes in the relationships within the household. Relationship changes that we know did occur include:

- a step-parent changing to adopted parent;
- an unrelated adult changing to step-parent; or
- a foster sibling changing to adopted sibling.

From the Wave 5 interview onwards the relationship of existing household members to the study child can be updated during the interview for household members present in previous waves.

As a result, from Wave 5 onwards there will be differences in the relationships between study children and household members between waves.

10.4.3 Inclusion of two waves of household data in the PLE person grid

The person grid is a list of people and their demographics associated with the study child, some members may still reside with the study child and others may have left. The Wave 5 parent living elsewhere survey instrument included roll-forward person grid data from Wave 4, so now two waves of household data for ongoing responding PLEs are available. Including Wave 4 details of a PLE's household in the survey instrument enables comparisons of the PLE's household circumstances between waves.

10.4.4 Concordance between people on main and PLE person grids

The concordance between the main household and the PLE's household has been provided for the first time in Wave 5. This enables the identification of who is the same person between the two files, who is on the main file only, and who is on the PLE file only. Table 18 provides a list of variables provided in the concordance file.

Table 18: Concordance file variables

Question ID	Label
MID5	Wave 5 Main Household Member Number
PLEID5	Wave 5 PLE Household Member Number
HHTYPE_5	Wave 5 Household Type
CHHFLOOP	Wave 5 Combined Household Row Number

The values for HHTYPE_5 are:

- 0 = Not present at Wave 5
- 1 = Wave 5 main household member only
- 2 = Wave 5 PLE household member only
- 3 = Wave 5 main and PLE household member

For example:

- Main household member number 4 was present at Wave 5, and that person was also present at Wave 5 in the PLE household, where they were recorded as member number 3. The variables that link these records will contain the following values: MID5 = 4; PLEID5 = 3, HHTYPE_5 = 3;
- If main household member number 4 was in the main household only at Wave 5, the values would be: MID5 = 4; PLEID5 = –9, HHTYPE_5 = 1;
- If PLE household member number 3 was in the PLE household only at Wave 5, the values would be: MID5 = –9; PLEID5 = 3, HHTYPE_5 = 2.

The values in MID5 and PLEID5 correspond to the member number in the data files, so this will enable you to find demographic information and link it to the files if required.

10.4.5 Child report of whether at school

At the start of both the study child's audio-computer-assisted selfinterview (ACASI) module and the face-to-face Child Self-Report K (CSRK) module, the interviewer records whether the study child is attending school, using response options of Yes and No. If the study child doesn't attend a school, some questions about schooling are not asked. These questions are directly related to the school environment and therefore are not relevant to study children not attending school. Parent 1 is also asked a question about whether the child:

- attends a government school
- attends a Catholic school
- attends an independent or private school
- is not in school.

In total, the number of K-cohort children coded as not in school as a result of the P1 interview was 33, whereas from the child interview the combined number was 218. Table 19 demonstrates that there were 191 records where the responses about whether the child was in school conflicted between the two interview components.

Table 19: Whether in school according to Parent 1 and study child components

	Study child (ACASI02/CSRK02)				
Parent 1 (EDUC14)	In school	Not at school (either question)	No study child interview	Neither question answered	Total
In school	3,639	189	51	38	3,917
Not at school	2	29	2	0	33
No P1 interview	4	0	0	0	4
Question not answered	1	0	1	0	2
Total	3,646	218	54	38	3,956

Table 20 cross-tabulates possible reasons for the discrepancy against school type, as recorded in the P1 interview for these 189 records. Around 44% of the difference seems to be accounted for by the interview taking place at the weekends or in school holidays.

Table 20: Characteristics of child or interview for children entered as not attending school by the interviewers

School attended	Interview date in school holidays	Interview date on weekend (not school holidays)	Interview date is school day	Total
Government school	32	16	61	109
Catholic school	11	5	23	39
Independent or private school	12	7	22	41
Total	55	28	106	189

To improve the quality of reporting in Wave 6, and to clear up any confusion, school attendance will be recorded in the same way in both the child interview and the Parent 1 interview. In the child interview the same response categories of government school, Catholic school, independent or private school, and not in school will be provided instead of Yes/No responses. This change is to make it clearer that the study is asking about usual school attendance and not whether school was attended on the current interview date. This point will also be further highlighted in interviewer training.

11 Smoking inside the household

In Wave 3 there was a higher number of families recorded as having 5 or more people who smoked inside the household than in other waves (see Tables 21 and 22 below).

Table 21: Number of residents who smoke inside – B cohort

B cohort	No. residents smoke inside									
	Wave 1	%	Wave 3	%	Wave 4	%	Wave 5	%	Wave 6	%
Refused (-3)	0	0.0	0	0.0	16	0.4	8	0.2	2	0.1
Not answered (-9)	766	15.0	0	0.0	184	4.3	251	6.1	217	5.8
Missing (.)	40	0.8	1	0.0	0	0.0	0	0.0	0	0.0
0	3815	74.7	4060	92.6	3759	88.6	3485	85.3	3276	87.0
1	318	6.2	136	3.1	205	4.8	228	5.6	174	4.6
2	141	2.8	58	1.3	68	1.6	97	2.4	86	2.3
3	18	0.4	3	0.1	6	0.1	13	0.3	6	0.2
4	5	0.1	1	0.0	1	0.0	1	0.0	1	0.0
5 or more	4	0.1	127	2.9	3	0.1	2	0.1	2	0.1
Total	5107	100.0	4386	100.0	4242	100.0	4085	100.0	3764	100

* This item was not collected in wave 2

Table 22: Number of residents who smoke inside – K cohort

K cohort	No. residents smoke inside									
	Wave 1	%	Wave 3	%	Wave 4	%	Wave 5	%	Wave 6	%
Refused (-3)	0	0.0	0	0.0	15	0.4	19	0.5	2	0.1
Not applicable (-9)	754	15.3	0	0.0	196	4.7	238	6.0	247	7.0
Missing (.)	54	1.1	1	0.0	0	0.0	0	0.0	4	0.1
0	3631	73.7	3939	90.9	3652	87.6	3366	85.1	2994	84.6
1	399	8.1	197	4.5	201	4.8	199	5.0	182	5.1
2	123	2.5	78	1.8	88	2.1	116	2.9	90	2.5
3	15	0.3	3	0.1	10	0.2	12	0.3	9	0.3
4	5	0.1	4	0.1	4	0.1	3	0.1	6	0.2
5 or more	2	0.0	110	2.5	3	0.1	3	0.1	3	0.1
Total	4983	100.0	4332	100.0	4169	100.0	3956	100.0	3537	100.0

* This item was not collected in wave 2

This difference in response is likely to be due to incorrect recording of 'none' responses as ‘5’ in the instrument as ‘5’ is the standard way for interviewers to record a 'no' response. In Wave 1 this item was collected as part of the Parent 1 leave behind form and in Wave 3 this question was collected by the interviewer in a face to face interview. However from Wave 4 onwards this question was changed to a Computer Assisted Self Interview (CASI) which the respondent completes themselves and as a result interviewer reporting error was not an issue. The change in collection mode may have also impacted the responses to other categories if there was response bias due to reporting smoking behaviours in a face-to-face interview rather than within the CASI which is completed alone by the respondent.

To correct this issue, responses to other waves and the number of people in the household were used to either amend the responses or set them to missing where it was unclear what the response should be.

Where reported responses to other waves was none, the Wave 3 data was set to none. If other responses were reported in other waves the data was set to missing, with the exception of 2 cases that reported 4 or 5 people smoking in other waves. This resulted in the following changes to the data shown in Tables 23 and 24 below.

Table 23: Wave 3 number of residents smoke inside amended results – B cohort

B cohort	No. residents smoke inside - chb15a4a			
	Wave 3 (original)	%	Wave 3 (amended)	%
Refused (-3)	0	0.0	0	0.0
Not applicable (-9)	0	0.0	0	0.0
Missing (.)	1	0.0	14	0.3
0	4060	92.6	4174	95.2
1	136	3.1	136	3.1
2	58	1.3	58	1.3
3	3	0.1	3	0.1
4	1	0.0	1	0.0
5 or more	127	2.9	0	0.0
Total	4386	100.0	4386	100.0

Table 24: Wave 3 number of residents smoke inside amended results – K cohort

K cohort	No. residents smoke inside - ehb15a4a			
	Wave 3 (original)	%	Wave 3 (amended)	%
Refused (-3)	0	0.0	0	0.0
Not applicable (-9)	0	0.0	0	0.0
Missing (.)	0	0.0	8	0.2
0	3939	90.9	4041	93.3
1	197	4.5	197	4.5
2	78	1.8	78	1.8
3	3	0.1	3	0.1
4	4	0.1	4	0.1
5 or more	110	2.5	0	0.0
Total	4331	100.0	4331	100.0

12 Missing data for Wave 6 items

12.1 Missing data for bullying items

In Wave 6 in the ACASI instrument B cohort children were asked ACASB 6.1:

During the last 12 months, since [month] last year…

a. kids hit or kicked me on purpose

b. kids grabbed or shoved me on purpose

c. kids threatened to hurt me

d. kids threatened to take my things

e. kids said mean things to me or called me names

f. kids tried to keep others from being my friend

g. kids did not let me join in what they were doing

h. kids used force to steal something from me

i. kids hurt me or tried to hurt me with a weapon

j. kids stole my things to be mean to me

k. kids forced me to do something I didn't want to do

l. I hit or kicked someone on purpose

m. I grabbed or shoved someone on purpose

n. I threatened to hurt someone

o. I threatened to take someone's things

p. I said mean things to someone or called someone names

q. I told others not to be someone's friend

r. I did not let someone join in what I was doing

s. I used force to steal something from someone

t. I hurt or tried to hurt someone with a weapon

u. I stole someone's things to be mean to them

v. I forced someone to do something they did not want to do

For selected categories children are then asked ACASB 6.5

How did this happen?

You can select more than one. (If you are not using a mouse, press space bar between responses.)

1. Face-to-face
2. Video chat
3. Phone call (not video chat)
4. Private messaging (includes email)
5. Open forum (e.g. Facebook walls, blogs, Twitter)
6. Other

Due to a sequencing error ACASB 6.5 was not asked for the following response categories of ACASB 6.1:

k. kids forced me to do something I didn't want to do

n. I threatened to hurt someone

o. I threatened to take someone's things

p. I said mean things to someone or called someone names

q. I told others not to be someone's friend

r. I did not let someone join in what I was doing

s. I used force to steal something from someone

v. I forced someone to do something they did not want to do

Thus, these items are not on the output file for the B cohort. The K cohort was not affected by this sequencing error.

12.2 Missing data for Cogstate items

In Wave 6, new tests or the K study children were given tasks to assess executive functioning. These tests are designed by CogState Ltd. For the LSAC, three tests have been included in instrument:

- Identification task (testing visual attention and choice reaction time) eg - press Yes if card is Red, No if card is Black.
- One back task (testing working memory) eg press Yes if current card is the same as the previous card, No if different.
- Groton Maze (spatial memory; impulse control and inhibition of erroneous responses) eg trying to work out the correct path through a maze and remembering it for the next time through.

For some of the Wave 6 interviews systems issues were encountered and as a result there are a small number of records with no Cogstate data on the output file. As Table 25 shows, the item hid40o1 indicates whether the Cogstate data is present and a reason for those records where the data is not present.

Table 25: Whether Cogstate data present – K cohort

14/15- EXF- Cogstate data present - hid40o1	Number of records	%
Not applicable (-9)	146	4.1
Cogstate data present	3234	91.4
Cogstate data not present - no consent given	43	1.2
Cogstate data not present - module could not be completed due to systems issues	83	2.4
Cogstate data not present - data loss due to systems issues	29	0.8
Cogstate data not present - child consented but did not complete any tasks	2	0.1

12.3 Missing data for puberty-related items

In Wave 6, all K cohort female study children were asked for the first time if they have ever menstruated (had your period). If yes, then the following questions applied about age of first period, any periods in last 3 months, and questions about menstrual problems in the last 3 months. These questions were asked as part of the Audio Computer Assisted Self Interview (ACASI).

In previous waves, the P1 was asked whether the study child had ever menstruated and the age of first period. These questions were asked as part of the Computer Assisted Self Interview (CASI) mode of the interview.

In Wave 6, a roll forward error in the instrument meant not all eligible children (those who reported at Wave 6 that they had started their period) were asked the subsequent questions about age started period and period experiences in the last 3 months. Those cases where the P1 had also previously reported that the study child's periods had started were not asked these additional questions.

Further, a sequencing error resulted in some study children (85) being asked if they had a period in last 3 months even though they answered no to ever had a period.

The Wave 6 data has been presented as collected from the study child. Items combining information collected from the two different informants have not been derived. Further, no data has been amended when answers between the two informants conflict or answers from the same informant conflict. For example, when P1 reported that the SC period had started, but the SC reported it hadn't and when the SC reported hadn't started period, but then reported one in last 3 months.

Table 26 below presents the various items that could be used for analysis of this topic. Only K cohort items have been presented given B cohort study children have not been asked these questions.

Table 26: Study Child menstruation items

Wave	Variable Name	Variable Label	Mode	Informant	Number missing
5	ghs36h	12/13 - CASI D4.7+W4 - Menstruate	CASI	Parent 1	NA
5	ghs36h1	12/13 - CASI D4.8+W4 - Menstruate (total months)	CASI	Parent 1	NA
6	hhs36h	14/15 - SC - ACASK 18.2 - Menstruate	ACASI	Study Child	NA
6	hhs36h1	14/15 - SC - ACASK 18.3 - Menstruate (total months)	ACASI	Study Child	876
6	hhs36i	14/15 - SC - ACASK 18.4 - Have you had any periods in the last 3 months	ACASI	Study Child	876
6	hhs36i1	14/15 - SC - ACASK 18.5.1 - How regular were your periods	ACASI	Study Child	876
6	hhs36i2	14/15 - SC - ACASK 18.5.2 - How heavy were your periods	ACASI	Study Child	876
6	hhs36i3	14/15 - SC - ACASK 18.5.3 - How painful were your periods	ACASI	Study Child	876
6	hhs36i4	14/15 - SC - ACASK 18.5.4 - How grumpy or teary did you get before your periods	ACASI	Study Child	876
6	hhs36i5	14/15 - SC - ACASK 18.6.1 - Did you miss any school days	ACASI	Study Child	876
6	hhs36i6	14/15 - SC - ACASK 18.6.2 - Did you miss any Social activities	ACASI	Study Child	876
6	hhs36i7	14/15 - SC - ACASK 18.6.3 - Did you miss any Sports or exercise	ACASI	Study Child	876

12.4 Missing data for Study Child helping others items

In Wave 6, K cohort study children were asked about helping others. More detailed questions about these situations were asked for up to 3 people. These questions were asked as part of the Audio Computer Assisted Self Interview (ACASI).

When study children entered something other than up to 3 names when asked who they helped (for example Paul John George Ringo, instead of just Paul John George), the instrument was unable to process who to ask the questions about and therefore these questions were not asked.

Table 27 below lists the items impacted and the counts. The second and third set of items have also been set to missing, given it is unclear how many people the study children may have helped.

Table 27: Amount of missing data for study child helping others items

Variable Name	Variable Label	Number missing
hsc28a	14/15 - ACASK 12.1 - Help someone with long-term health condition/disability/elderly	NA
hsc28b1	14/15 - ACASK 12.3.1 - What is first persons relationship to you	132
hsc28b2	14/15 - ACASK 12.4.1 - Does first person live with you	132
hsc28b3	14/15 - ACASK 12.5.1 - Does first person go to the same school as you	132
hsc28e1	14/15 - ACASK 12.6.1.1 - Help provided - 1st person - Personal care	132
hsc28f1	14/15 - ACASK 12.6.1.2 - Help provided - 1st person - Moving around	132
hsc28g1	14/15 - ACASK 12.6.1.3 - Help provided - 1st person - Transport	132
hsc28h1	14/15 - ACASK 12.6.1.4 - Help provided - 1st person - Communicating	132
hsc28i1	14/15 - ACASK 12.6.1.5 - Help provided - 1st person - Preparation of meals	132
hsc28j1	14/15 - ACASK 12.6.1.6 - Help provided - 1st person - Housework/shopping/errands	132
hsc28k1	14/15 - ACASK 12.6.1.7 - Help provided - 1st person - House repairs/garden care	132
hsc28l1	14/15 - ACASK 12.6.1.8 - Help provided - 1st person - Health care	132
hsc28m1	14/15 - ACASK 12.6.1.9 - Help provided - 1st person - Paperwork	132
hsc28n1	14/15 - ACASK 12.6.1.10 - Help provided -1st person - Keeping them company	132
hsc28o1	14/15 - ACASK 12.6.1.11 - Help provided -1st person - Other	132
hsc28q	14/15 - ACASK 12.8 - How often do you do these caring activities	Not impacted as not in a loop
hsc28r	14/15 - ACASK 12.9 - On average - total number of hours you spend providing care	Not impacted as not in a loop

13 References

Australian Bureau of Statistics. (2006). *Australian and New Zealand Standard Classification of Occupations*, First Edition, Cat no. 1220.0. Canberra: ABS.

Australian Bureau of Statistics. (2002). *Australian Standard Geographical Classification (ASGC) 2002*. ABS cat. no. 1216.0. Canberra: ABS.

Australian Bureau of Statistics. (2011). *Australian Statistical Geography Standard (ASGS): Volume 1. Main structure and greater capital city statistical areas*. Cat .no. 1270.0.55.001. Canberra: ABS.

de Lemos, M. & Doig, B. (2000) *Who Am I? Supplementary Information.* Melbourne: ACER.

Dunn, L. M. & Dunn, L. M. (1997). *Peabody picture vocabulary test*, 3rd edition. Circle Pines, MN: American Guidance Service.

Egerton, M. & Gershuny, J. (2004). *Utility of Time Use Data: Report to DfES*. ISER University of Essex.

Fisher, K. (2002). *Chewing the fat: the story time diaries tell about physical activity in the United Kingdom*: Institute for Social and Economic Research, University of Essex.

McMillan, J., Beavis, A., & Jones, F. L. (2009). The AUSEI06: A new socioeconomic index for Australia. *Journal of Sociology*, *45*(2), 123–149.

Rothman, S. (June 2003). *An Australian version of the Adaptive PPVT-III for use in research*. Unpublished paper. Melbourne: Australian Council for Educational Research.

Wechsler, D. (2004). *The Wechsler intelligence scale for children*, 4th edition. London: Pearson Assessment.

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