Growing Up in Australia

The Longitudinal Study of Australian Children:

an Australian Government initiative

LSAC Discussion Paper No. 5

Wave 2 Data Management Issues

Sebastian Misson

June 2007

Australian Government
Australian Institute of
Family Studies

Contents

Introduction

This paper presents a discussion of the data management policy and procedures for Wave 2 of *Growing Up in Australia* - the Longitudinal Study of Australian Children (LSAC).[1]

This paper discusses key data management issues associated with the project that are pertinent to Wave 2:

- Variable naming conventions;
- File structure;
- Treatment of household composition data;
- Data confidentialisation;
- Data imputation; and
- Weighting of data.

This paper reviews current policy and procedure as implemented following the release of Wave 1 data and further discusses some potential revisions and additions to these in light of the forthcoming release of Wave 2 data, which will mark for the study the first major release of a longitudinal dataset. An overview of the Data Management Principles that guide LSAC can be found in *LSAC Discussion Paper No. 3: Data Management Issues*. Further discussion of issues relevant to Wave 1 data management can be found in the *LSAC Data Users Guide*. Interested parties wishing to provide comment, should do so by emailing Sebastian.Misson@aifs.gov.au before the July 6 2007.

[1] *Growing Up in Australia* was initiated and funded as part of the Australian Government's *Stronger Families and Communities Strategy* by the Australian Government Department of Families, Community Services and Indigenous Affairs. The study is being undertaken in partnership with the Australian Institute of Family Studies, with advice being provided by a consortium of leading researchers at research institutions and universities throughout Australia. The data collection is undertaken for the Institute by the Australian Bureau of Statistics.

Variable Naming Conventions

The Wave 1 variable naming scheme for questionnaire items was based largely on the questionnaire positioning of the question that produced the variable (e.g. b1cb34 was the B-cohort, Wave 1 item corresponding to question B34 of the face-to-face interview). This has some key implications as we look towards having multiple waves of data collected largely by CAI:

- Adding, removing and changing the order of questions or moving them to different sections or questionnaires means that some variables could become 'lost', or at the very least that users would require a lot of reference material to navigate the data set. For example, the question assessing a child's general health is number 34 in the health section for the B-cohort in Wave 1, number 23 for the K-cohort in Wave 1, and number 19 for both cohorts in Wave 2. By the time Wave 4 comes around the variable might have moved to a different position two more times. This would mean a lot of work for users to find it in every wave before conducting a longitudinal analysis.
- Question numbers in the CAI aren't quite so straight forward as in paper questionnaires, meaning that numbering can become more complex and use greater numbers of characters (eg. if you want to squeeze a new question in between Q1a and Q1b, it becomes Q1a1 and so forth). This could lead to the need for unpredictable variable names as older versions of SPSS impose a limit of 8 characters per name.

A new variable naming convention is therefore being proposed. The implication of this is that all Wave 1 variable names will need to be changed, however data users will be provided with code to change Wave 1 variables back to their original names if they have code written for the old names which they still wish to use.

Some guiding principles in the construction of new names were that variable names should:

- be consistent across cohorts for easier merging of files;
- be predictable across waves to reduce the need to look up variable names;
- be no more than 8 characters long to enable use in less recent versions of SPSS; and
- use as few characters as possible to allow greater flexibility for future Waves.

In the interest of satisfying the last criterion, removal of some information from the variable name is proposed. Firstly, including an indicator of survey instrument leads to unnecessary complications if items move from one instrument to another.

Secondly, including a cohort indicator diminishes the user's ability to merge variables containing identical information across cohorts as variable names of identical information will be different. For example, in Wave 1 k1ca5sc was the study child's relationship to P1 for the k-cohort, while b1ca5sc contained this information for the b-cohort. To run an analysis using this data for both cohorts involved either coding two separate analyses or renaming one or both the variables and merging the datasets.

A further proposed improvement would be to switch from using a wave indicator to a child's age indicator. Keeping a wave indicator on the data means that that cross-sectional analyses involving the latest waves of data are facilitated (e.g. how many children have separated parents at Wave 2 of data collection?). However, as the study's dataset becomes more longitudinal, it is anticipated that comparing the children at like ages is going to be more prevalent. (e.g. At what age are the parents of a child most likely to separate?). Using an age indicator make it easier to use both cohorts' data to answer these questions, and test for difference in the 4 years between cohorts.

Given these considerations, it is proposed that variable names take on the following form:

A tt xxxxx

Where:

A: Child age indicator.

tt: topic indicator

xxxxx: specific question identifier.

Child age indicator (alpha)

This will link variables together across cohorts when the child is the same age, e.g. for the B-cohort the letter would be *a* for Wave 1, *b* for Wave 2, *c* for Wave 3, while for the K-cohort it would be *c* for Wave 1, *d* for Wave 2, *e* for Wave 3 etc. Those items of information that are permanent once decided (e.g. details of birth, age began or stopped something, etc.) will be given the age indicator z so that they will have consistent name across cohort regardless of the age of the child when the information was obtained.

Topic indicator (alpha)

Taken from the topic field of data dictionary (although some alterations may be useful), abbreviations will be meaningful (e.g. family demographics will be *fd*, child's development will be *cd*). Some revisions to the topic indicators used in Wave 1 might be required to enhance their usability for this purpose, while others will need to be added as the survey instruments evolve.

Specific question identifier (alphanumeric)

These last 5 digits will contain whatever information is necessary to uniquely identify each item. Each will obtain an arbitrary question number, not related to their questionnaire positioning. Items will be grouped together into questions as much as possible. For example, all items that form a scale will have a single question number and, to the extent that this is possible while staying under 8 characters, will also identify layers of structure within a question (e.g. sub-scales and sub-sub-scales). The informant/subject of the question will also be identified in these digits.

The following examples from Wave 1 show how the naming convention will work in practice (however all new names should be considered strictly as drafts):

- Birth weight of the study child: Currently *b1cb4* and *k1cb2* for the two cohorts, it will become zcp4 for both cohorts, ie *z* as the age indicator as it will not change with time and *cp* for the topic 'conception, pregnancy and birth' (the 4 is arbitrary based on it being the 4th question allocated a number).
- Parent rating of parenting self-efficacy: Currently *b1cf1* for Parent 1 and *b1sa16* for Parent 2 for the B-cohort, and *k1cf1* for Parent 1 and *k1sa9* for Parent 2 in the K-cohort. This will become *apa1a* for Parent 1 and *apa1b* for Parent 2 for the infant cohort, and *cpa1a* for Parent 1 and *cpa1b* for Parent 2 for the K-cohort. Note that for the infant cohort the mother v father versions of these variables will predictably be called *apa1m* for mothers and *apa1f* for fathers, and that if the

item is still in use in Wave 3 the B-cohort items will have the K-cohort names from Wave 1 (ie *cpa1a* and *cpa1b*).

- Parent 1 parenting hostility scale: Currently *b1cf12* to *16* for the infants and *k1cf11, 12, 13, 17* and *18* for the 4-5 year olds. Because these are actually different scales they will be given separate question numbers, so the B-cohort version will be *apa4aa* to *apa4ea*, and the K-cohort version will be cpa11aa to cpa11ea.
- The Strengths and Difficulties Questionnaire, prosocial subscale: This is a scale that both Parent 1 and the teacher fill out. It is currently it is *k1cf21, 24, 29, 37* and *40* for the Parent 1 version and *k1tc10, 13, 18, 26* and *29* for the Teacher version. Under then new scheme it will become *cdv8a1a* to *cdv8a5a* for the Parent 1 version and *cdv8a1t* to *cdv8a5t* for the teacher version. Note that the *8a* is there because it is a subscale of a larger scale; so for example the hyperactivity subscale items will be *8b*, the emotional symptoms subscale items will have *8c*, etc.

The above proposal has received wide distribution among stakeholders since being developed in October 2006. Their feedback has generally been positive, with only one minor change to the proposal being adopted. This change was that 'permanent' items, such as date of birth, receive a nominal age indicator *z*, instead of having no age indicator at all. This was to allow greater compatibility with the PanelWhiz package under consideration for use with the LSAC dataset in STATA. While the above proposal has received widespread support, there are some parts that involved trade-offs between competing priorities that should be highlighted:

- Wave naming: It is proposed to switch from wave naming of variables to age naming variables as described above. Wave naming refers to the system used for Wave 1, where the Wave number was part of the variable's name. In this system the two cohorts can more easily merge data from the same wave, however merging data when children are at the same Wave is slightly more problematic. Age naming is recommended over wave naming is felt that the most valuable analyses looking into the future of the study are likely to be longitudinal. There is also likely to be more common questions when cohorts are identically aged compared with when the children are at different ages, but data collection is simultaneous.
- Dealing with teachers/carers as informants: Depending on the age of the child, at each wave of data collection each case can include data from a teacher, centre-based or home-based carer questionnaire, but not from more than one of these. At Wave 1, each of these received a separate letter to distinguish their variable name (i.e. home-based carer items were coded *g*, centre-based carers coded *l*, and teachers coded *t*). In order to allow for greater consistency, it is proposed that a single informant code be used for all of three of these groups. Therefore, carers will have their data merged into the same variables when the questions are identical and teachers will continue to use the same names for identical questions when they take over care of the children (age indicators aside). A variable will be provided so users can distinguish which cases had centre-based carer questionnaire data, and which had home-based carer questionnaire data at each of the 1st two waves.
- Naming of derived items: Under the proposed scheme derived items could be named in relation to their component items (e.g. in the scale formed by their

mean of items Q3a, b and c could be called Q3) rather than the current situation where the derived item names are meaningful. However, this is not recommended as it is felt that the value of the transparency of knowing which items went into each derived item is less than the mnemonic assistance provided by meaningful names. It is not recommended to move questionnaire items to more meaningful names as the number of items makes naming conflicts and other sources of confusion hard to avoid.

File Structure

Wave 1 Data Structure

The main data file for each cohort in Wave 1 contained approximately 2,000 variables. The following represents the basic outline of how these variables were ordered in the file:

- Identifiers and status variables (e.g. form response flags)
- Questionnaire items:
    - Face-to-face interview
    - Parent 1 Self-complete
    - Parent 2 Self-complete
    - Home-based Carer questionnaire (B-cohort only)
    - Centre-based carer questionnaire (B-cohort only)
    - Teacher questionnaire (K-cohort only)
- Derived items
- Weighting variables
- Neighbourhood characteristics (e.g. location, linked census data)
- NCAC linked data
- Mother/Father variables (ie questions asked of P1 and P2 separately reframed so all mothers go together and all fathers go together)
- Wave 1.5 data

In addition to this main dataset, the Time Use Diaries and the Medicare Australia linked data were provided in separate datasets as these were best presented as files with multiple records per child. These datasets could be linked back in with the main dataset (and each other) by a shared child id number.

Consideration needs to be given also to how many data files should be used to contain the data when Wave 2 is added. Having one large data file requires less merging of datasets, but often slows processing speeds when using the data as the computer has to navigate amongst much extraneous information to get the variables it needs. Having too many small data sets requires much merging of files and customisation of datasets so is also not ideal. It is proposed that each wave of data be given its own dataset which will contain all information currently in the main dataset. Supplementary datasets for time-use diaries and Medicare data will also be provided for each wave.

With Wave 2 data becoming available some different options can be considered in particular with regard to the main datasets.

One of the issues data users have had with the LSAC data is finding the derived items that correspond with questionnaire items. While there are a number of ways users can do this easily via documentation such as the data dictionary and the labelled questionnaires, making the links more obvious within the dataset may be beneficial. It is therefore recommended that the derived items and the questionnaire items are interleaved so that derived items are next to the items they are derived from. For example, in the Wave 1 data set the items that make up the Parent 1 SDQ are k1cf21 to k1cf45, which come after k1cf20 and before k1cg1. Later in the dataset, in the derived items section, are the SDQ derived items ap1psoc, ap1hypr, ap1emot, ap1cond, ap1peer and ap1sdqt. The preferred alternative is to have the SDQ derived items come after k1cf20 and before k1cf21.

Another way of presenting the data would be to group items in the same topic group together. This would give the opportunity for equivalent variables from different waves to be placed next to each other with each topic presented in it's own dataset. Such a presentation could be a large advantage in setting up longitudinal analyses when using graphical interfaces, but might require too many small datasets. Also, correspondence between the order of variables in the questionnaire or CAI instrument and order of variables in the dataset would be lost. Additionally, the classification of variables would need to be intuitive, which is likely to be difficult for those variables which could fall into two or more subject groupings. Given these complexities it is recommended that this not be attempted.

Household composition

The data collected on household composition is designed to represent a continual picture of the comings and goings from the study child's home, rather than a snapshot at just the time of data collection. For this reason, and due to the large amount of data that is simply transferred from Wave 1, it does not make sense to treat the household composition module as a separate set of questions at each wave, but rather to merge the household information from multiple waves into a single set of variables. This should not prevent users from taking a snapshot of household composition at a time, but should rather facilitate other analyses focussed on change by restricting the number of variables required to store the information as much as possible.

The current proposal for the structure of the household information is to have a grid of variables for every member of the home. For every household, the Study Child will be Member 1, Wave 1's Parent 1 will be Member 2, and Wave 1's Parent 2 will be Member 3. Any additional people in the household at the time of Wave 1 will be given Member numbers 4 through to whatever is required. When a new person is found in a home at a particular wave they are given the first number that has not been used by anyone previously for that household and their details are loaded into the appropriate positions in the grid. Users will know which waves each member was present and absent for via flag variables that will have this information. It should be noted that numbers are never re-assigned, even if a person leaves the home, and that this has a few implications.

Firstly, even though we always know that Wave 1's Parent 1 will be Member 2, there will be no set position for Wave 2's Parent 1. In the vast majority of cases Member 2 will still be the Parent 1 at Wave 2, however it is possible that the Wave 2 Parent 1 could be Wave 1's Parent 2, or someone else who was in the household at Wave 1, or even someone new to the household in Wave 2. In order to deal with this confusion there will be a variable at each wave specifying the member number of Parent 1 and Parent 2. Additionally, two new sets of variables will be created with the characteristics of the Parent 1 and Parent 2 for each wave (ie if Member 2 is still Parent 1 these variables will be loaded with Member 1's details, if Member 5 is now Parent 1, these variables will be loaded with Member 5's details, etc.). Flag variables will also be set up at each wave so that users can tell whether Parent 1 and Parent 2 have changed since the last wave.

Secondly, even if there was no Parent 2 at Wave 1, 3 will never be assigned as a member number to any other person. This means that the Member 3 set of variables will always just have the details for Wave 1’s Parent 2 without other cases having to be filtered out.

Thirdly, the distinction between ‘parents’, ‘siblings’ and ‘others’ that existed in the household in Wave 1 will be removed as all just become members (although Parent information will be recorded separately as well, see above). This is due to the need for each member to have a consistent position in the dataset throughout. For example, there will be situations where a grandparent starts off as a parent, but then may become an ‘other’ as the study child’s biological parent takes over more responsibility, even though they stay in the household.

Confidentialisation

In Wave 1, data files have been released with two different levels of confidentilisation:

- De-identified data, and
- Moderately confidentialised data.

De-identified file

In this file, all name, address and other contact details have been removed for the child, family, childcare agency and teacher or carer. The main datasets contain identification numbers for each child, which can only be linked to specific personal details by the fieldwork agency. No released datasets contain this personal information.

The de-identified data file is an output file with only the above information removed.

Moderately confidentialised file

In Wave 1, the moderately confidentialised file was created by performing the following alterations:

- All names and contact details removed
- Qualitative data provided by respondents removed
- Postcodes were given an indicator so that all children selected in the same postcode can be identified
- Date of birth transformed to age in months at time of interview and month of birth
- Date left hospital after birth derived as number of days between birth and departure
- Parents' occupation aggregated to 2-digit ASCO level
- Occupation in previous job aggregated to 2-digit ASCO level
- Income top-coded
- Housing costs top-coded
- Child support paid by parent 2 top-coded
- Children's current height, weight and waist circumference measurements - top-coded
- Number of hours spent in childcare top-coded
- SEIFA variables rounded to the nearest 10
- Country of birth recoded to 0 if fewer than five others have same code for variable
- Religion recoded to 0 if fewer than five others have same code for variable
- LOTE recoded to 0 if fewer than five others have same code for variable

The Wave 1 versions of all these items will remain confidentialised in the Wave 2 dataset, and if collected again at Wave 2, these items will receive the same confidentialisation. It is not currently considered necessary to confidentialise any of the items that are entirely new to Wave 2, however this decision is still under review.

Data Imputation

Data imputation has been used in LSAC at Wave 1 only to improve data quality for items where problems were known to exist (e.g. k1cc34, time-use diaries). Most commonly, any issues such as these that arise in Wave 2 will need to be responded to on a case-by-case basis as they come to light. However, it is already possible to predict some areas that might require attention.

Virtual roll-forward

"Roll-forward" is the term in CAI design that refers to the use of data from a previous wave of data collection to determine the questions that need to be asked in a subsequent wave. For Wave 2 a limited set of data was rolled forward, largely to assist with the household composition module. Time and resource implications meant that this could not be implemented in some other parts of the questionnaire where it may have reduced respondent burden. For example, in Wave 2 we re-ask respondents about the age child stopped being breastfed in order to obtain the information from those cases where this had not yet happened at the time of Wave 1. In re-asking this questions it is likely that some respondents will give different answers from their Wave 1 responses. Given the recollection of respondents is likely to be more accurate closer to the event (ie the cessation of breastfeeding), it is proposed that in cases where Wave 1 data exists the Wave 1 value is taken as correct, and the Wave 2 value is ignored (ie as if the Wave 1 data was rolled forward and the question was never asked in Wave 2). From the data users perspective, a single variable is produced that represents the best estimate from the two waves of data (ie as if roll-forward had occurred). Users will be able to tell at which wave the timing data was collected by the question from each wave asking if the child is still being breastfed.

Longitudinal contradictions

Another possible error from the above is the situation where respondents report at Wave 2 that an event (again breastfeeding cessation is an example) occurred at a time before Wave 1, when Wave 1 data indicates this event hadn't happened yet at that time. In this case it is proposed that the time of Wave 1 interview be treated as the time of the event. For example, if a parent reported at Wave 2 that the child stopped being breastfed after two months, however at Wave 1 the child was 3 months old and was reported as still being breastfed, the age of breastfeeding cessation would be set to 3. Other examples where this rule would apply include time of entry for new people in the home and length of attendance at different childcare types. Obviously if there is any case where this rule would not be logical, the imputation will not be made. A full list of variables where this rule will be used will be approved by FaCSIA and provided in the data user guide.

Time-use diary

While steps have been taken to ensure that the problems that existed in Wave 1 with false-positives will not re-occur in Wave 2, some imputation of the time-use diary will be made to improve data quality. Wherever possible this imputation will be based on the Wave 1 rules to avoid biasing estimates of change between waves. As per Wave 1, some cases where the data is still of such poor quality that its inclusion might provide more harm than good for the analysis will be deleted from the cleaned file. Again, extra files will be provided to users with the complete raw data, and the cleaned data for any cases that were deleted from the final TUD file. This will ensure that users can check the effect of these imputations and deletions.

Weighting

Weights in the LSAC data set in Wave 1 were used to provide some measure of correction for biases in the sample design and non-response of potential respondents. The final weights put on the file were based on design weights, calculated from the inverse of the chance of selection to be invited to participate in the study. These design weights were then adjusted to correct for the most important sources of non-response bias that could be identified, the mother's educational level, and the mother's use of a language other than English at home.

Two weights were published on the data file as a result of these calculations:

- A population weight that adjusted estimates of frequencies produced by the data to population totals (e.g. x number of children in Australia had characteristic y)
- A sample weight that adjusted estimates of percentages produced by the data to the proportions given when using the population weight, but kept the frequency estimates reflective of the number of children in the sample (e.g. x number of children in the LSAC sample had characteristic y). This second weight should be used when tests of significance are to be generated.

While it would have been possible to provide separate weights to adjust for forms non-response (e.g. to adjust for non-response bias in estimates produced by the Parent 1 Self-Complete Questionnaire (P1SC)), this was not attempted. It was considered this would add an extra amount of complexity as it would require users to select between a number of weights. While this might not be difficult for experienced data analysts, it is intended that the LSAC data be easily used by people with different skill sets in order to increase the value of the study to the community. The selection of weights might be more complicated when data from two different forms were being used in the one analysis, (e.g. if crosstabbing a result from the P1SC with one from the Teacher Questionnaire, should the P1SC weight be used or the Teacher Questionnaire one?). It could be argued that it would be necessary to produce weights to deal with every combination of questionnaire non-response, however looking longitudinally this would quickly become extremely complex and resource intensive to generate.

In considering whether to update the weights at Wave 2, a similar trade-off between complexity and accuracy needs to be made. Weighting does provide some correction for non-response bias, but some reasons for non-response will remain hidden from any potential for correction. It is important that the limitations of weighting be understood so that such a trade-off can be assessed.

The primary problem in correcting non-response in LSAC is that the biggest source of bias comes from those families that were selected to participate but did not become part of the Wave 1 dataset, either due to difficulties making contact or refusal to participate. Around 20,000 families were invited to participate in LSAC, of which a little over 10,000 did. The erosion from 10,000 families to 9,000 in Wave 2 represents much less of an opportunity for the introduction of bias. This is problematic since very little can be known about people who never participated in the study and how their participation might affect the estimates produced by the study. The weights produced in Wave 1 identified important factors in non-response by looking at the effect that the characteristics of a postcode's population (eg ethnicity, financial situation etc.) had on the response rate for that postcode. This means that investigation of this initial non-response was limited to variables that were measured by both the Census and LSAC, with the implication that variables that contributed to non-response that did not meet these criteria could not be adjusted for.

More will be known about those that do not respond in Wave 2 from their data at Wave 1. However, this data will be based on their lives two years ago. In the intervening time a number of major changes may have happened in the lives of the LSAC families, such changes include those of location, employment, family structure, financial situation and health status. Non-response can happen for any number of reasons and would be difficult to predict accurately with contemporaneous data, therefore any adjustment can only hope to account for a modest proportion of the bias created.

A final issue to confront weighting is that the use of LSAC for cross-sectional estimates will necessarily involve greater amounts of error over time. The LSAC population represents a group of children that were resident in Australia at the time of sample selections. It will not include any children that have moved to Australia since May 2004. For example, this means that very few children in the B-cohort are born overseas.

From the preceding discussion it should be clear that any weighted dataset would not be free of non-response biases, so too much complexity in the production of weights would seem not to be warranted. However, even given the limitations outlined, the introduction of a new weight for the Wave 2 interview will present some measure of correction for on-going non-response without introducing too much complexity to the data.

It is anticipated that the weight would be calculated by the following process.

- Run a logistic regression to estimate the probability of each family from Wave 1 completing the interview in Wave 2.
- Divide each case's Wave 1 weight by this probability for all cases that had responded to Wave 2 (so that high probability cases have relatively lower weight and low probability cases have relatively higher weight).
- Adjust total weights for each strata so that the proportion for each selection stratum is what it was following Wave 1 weighting.
- (If necessary) Topcode and bottom code extreme weights and recalibrate stratum to have correct proportions.
- Adjust all weights so that average values are appropriate, ie mean value of 1 for the sample weights, mean value of (population size/sample size) for population weights.

This approach to adjusting initial weights for non-response using logistic regression is similar to those used in other longitudinal studies such as the Household Income and Labour Dynamics in Australia Survey (Watson, 2004), the Panel Study of Income Dynamics in the US (Gouskova, 2001), and to a slightly lesser extent the National Longitudinal Study of Children and Youth in Canada (Statistics Canada, 2006).

Dataset Version Naming

Following the release of the Wave 1.5 data, the version number of the dataset was set to version 2.0. Some users have been concerned that this would lead to confusion when Wave 2 data was released. Furthermore, other users had trouble discerning if the data they received had certain updates included on them. To resolve these issues future datasets will be issued with a Wave indicator and a time indicator. So, for example, the initial release of data for Wave 2 will not be Wave 2 v1.0, but rather Wave 2 August 2007.

Conclusion

This paper has highlighted some of the issues that need to be considered in managing LSAC as a longitudinal data set. In summary, it is recommended that:

- variable names be revised for greater consistency between waves and between informants;
- each wave of data should be kept in a separate dataset with separate supplementary datasets at each wave for time use diary data, and Medicare Australia linked data;
- derived items be placed in the dataset near the questionnaire items that they were derived from, rather than as a single block after all the questionnaire items.
- teacher-carer data be into a single set of variables under the same informant indicator, rather than using three separate ones as is the case now.
- confidentialisation should follow the same procedures as laid down in Wave 1;
- data collected at Wave 1 be used in preference to that collected in Wave 2 where contradictions between the two exist;
- a process of imputation of time use diary data be carried out which closely mimics that performed after Wave 1;
- additional weights be calculated to adjust for non-response bias to the Wave 2 interview;
- further weights to adjust to non-response for forms, not be calculated; and
- version naming of datasets to include the wave and month of release.

It is hoped that these recommendations will mean that the LSAC data continues to be user-friendly and accurate, encouraging its maximum utilisation.

References

Gouskova, E. (2002). *The 2002 PSID Child Development Supplement (CDS-II) Weights.* PSID Technical Report.

Statistics Canada (2006). *National Longitudinal Survey of Children and Youth, Cycle 6 – User Guide*. Ottowa, Canada.

Watson, N. (2004). *Wave 2 Weighting*. HILDA Project Technical Paper Series No4./04. The Melbourne Institute of Applied Economic and Social Research, The University of Melbourne.