The Longitudinal Study of Australian Children

Data Management Issues

Discussion Paper Number 3

Robert Johnstone,
the Project Operations Team and
the LSAC Research Consortium[i]

March 2004

LSAC was initiated, and is funded by, the Australian Government Department of Family and Community Services

Contents

Introduction

This paper presents a discussion of the data management policy and procedures for *Growing Up in Australia* - the Longitudinal Study of Australian Children (LSAC). *Growing Up in Australia* is a major study funded by the Australian Government Department of Family and Community Services (FaCS) as part of the Australian Government's *Stronger Families and Communities Strategy*. The Australian Institute of Family Studies is leading a consortium of nine eminent Australian research institutions in the development of this study, which will track the development of two cohorts of young children for at least 7 years.

Growing Up in Australia is one of the largest and most complex studies of this nature that has ever been undertaken in Australia. The study aims to provide the data for a comprehensive understanding of Australian children's development in the current and future social, economic and cultural environment, and hence to become a major element of the evidence base for policy and practice regarding children and their families. *Growing Up in Australia* is now also part of the Government's move towards the development of a national agenda for early childhood.

This paper discusses key data management issues associated with the project; namely:

- Data management principles
- File structure
- Data confidentialisation and access
- Data security
- Data imputation
- Outliers
- Weighting of data
- Respondent tracking
- Data linkage
- Software implementation

This paper presents principles to guide policy and procedure, and in many cases oulines our currently preferred options. During the months before receipt of Wave 1 data these will be further considered and refined. We welcome comment from readers, including future users of the data and other stakeholders. These comments will be taken into consideration when finalising the approaches to be taken.

In taking a comprehensive approach to each of these issues, the paper outlines best practice to ensure the integrity of data management of *Growing Up in Australia* for the duration of the project. Issues such as imputation, confidentialisation and weighting will be further discussed in future technical papers.

Some of the paper content is drawn from similar publications – in particular the University of Melbourne's Household, Income and Labor Dynamics in Australia (HILDA) Project Discussion Papers 1/01, 2/01, 3/01 and 3/02. For some issues, this longitudinal study has already established procedures that are appropriate for LSAC. In other instances differences in study design limit the extent to which common strategies can be adopted in these two longitudinal projects.

Data Management Principles

Data management embraces the whole range of activities involved in the handling of data. These activities include: data related policy; data ownership and responsibilities for ensuring legislative compliance; data documentation and metadata compilation; data quality and standardisation; and data access and dissemination.

The LSAC data is owned by the Australian Government Department of Family and Community Services and governed by the Privacy Act, the Health Insurance Act, the Crimes Act, the Commonwealth Protective Security Manual, the Australian Communications – Electronic Security Instruction and the National Health Act. The Australian Government Department of Family and Community Services has contracted AIFS to undertake most data management tasks of the LSAC project.

The LSAC study design aims to ensure the collection of high quality data. The work of data management complements this by checking the validity and consistency of the data and ensuring that the data can be readily used by researchers to best answer the myriad of research questions about children's early development.

The LSAC data management team will be part of the AIFS project operations team. The goals of the LSAC data management team are to:

i. Produce quality data;
ii. Develop a file structure and coding framework that allow for ease of use of the dataset;
iii. Assure that data received from the field agency meets stringent levels of quality assurance;
iv. Ensure the adoption of methods which reflect best practice in relation to imputation, weighting and confidentialisation;
v. Provide consistent and acceptable documentation;
vi. Deliver services to researchers (e.g. training); and
vii. Implement the FaCS LSAC data access policy

The prime concern of the data management team will be to ensure high quality, ease of access to, and use of, data by researchers while protecting respondent privacy. This requires that data be systematically and logically organised.

File Structure

Principles of Proposed Data Structure

The LSAC dataset, in its rawest form, will contain over 17,000 individual data items, each related in some way to the study child. A means for ordering this wealth of data and turning it into information for the end user is essential. Efficient and effective delineation of the dataset(s) is required to ensure researchers thoroughly mine the wealth of resources available within the LSAC datasets.

File Format

The survey data will consist of a series of cross-sectional datasets (linked to enable longitudinal anlysis) for each year of collection with the child as the reference point. These datasets will allow simple cross-sectional analyses to be conducted.

Two options for data storage are being considered. Option (a) would present data as a single flat file for each cohort where records have the same number of "fields". There would be a record for each individual child and the child's unique identification number would serve as the key.

Option (b) would create a series of cross sectional datasets for each cohort, delineated by conceptual area (e.g. health, education and child care). Connecting different datasets would be achieved by using a common link or key. With this common link datasets may be combined by the user to form new inquiries and data output.

Option (a) provides a simple method of data storage. Data would be accessed within a single dataset. Thus analysis across thematic areas would not require any data linkage. However, due to the number of data items collected, end users may find current computer hardware to be underpowered, limiting analysis speed, and thus the usability, of the final, sizeable dataset. Option (b) would allow speedier access and analysis of thematic subsets of the complete LSAC dataset. Analysis across themes would be more complex.

Regardless of whether option (a) or (b) is pursued, the file structure of the LSAC datasets needs to be relatively intuitive in its logic, thus allowing easy identification of variables for analysis and to enable longitudinal analysis. For both option (a) and (b), it is proposed that data be stored according to the order shown in the collection instrument and be thematically cross-referenced in the LSAC data dictionary. To complement this ordering, the questionnaires will be made available to users with variable names specified against each data item, allowing for efficient, more accurate identification of data items of interest. Furthermore, the data dictionary will allow users to access the LSAC metadata (information about the data), search for material according to instrument, thematic grouping, variable name and label and download this material to assist data analysis.

For the purposes of understanding the dataset layout, a brief summary of the study instruments is presented.

Overview of the Survey Instruments

The first wave of LSAC involves the use of 13 questionnaires and other instruments. The large number of instruments is due to:

(i) The design of LSAC is cross-sequential, meaning that two cohorts of children, aged 0 and 4 years, will be followed over time. While much common content exists across cohorts, separate questionnaires with age appropriate items for each cohort have been developed.

(ii) The unit of interest in LSAC is the child, and the informants include the child's resident parents, in-home and centre-based child care providers, and teachers. Further data is derived from observations by interviewers during the period they are in the child's home and neighbourhood. Direct assessment of physical and cognitive development is also undertaken.

For each cohort interview, some data are best collected through face-to-face interview and some by self-completion questionnaire. The first wave of LSAC will include the following instruments:

(i) the Parent 1 Face-to-Face;

(ii) the Parent 1 Self-Complete Questionnaire;

(iii) the Parent 2 Self-Complete Questionnaire;

(iv) the Time Use Diary; and

(v) the Family Contact Form;

In addition to these instruments, for the infant cohort there is a Home-based Carer Questionnaire and a Centre-based Carer Questionnaire, while for the 4 year old cohort there is a Teacher's Questionnaire, the Peabody Picture Vocabulary Test (PPVT) and the "Who am I?" (WAI) assessment.

All output will be in relation to the child. Various outcome data will be collected, including factors influencing these outcomes. This will allow the investigation of key research questions (e.g. how are child outcomes related to particular aspects of the child care they receive). More detailed information on these issues and the study design can be found in the LSAC discussion papers available from the AIFS website.

The rest of this section will give an overview of each instrument used in the first wave of data collection.

Parent 1 Interview (P1N)

The P1N is the main instrument for data collection. For the purpose of the study Parent 1 is defined as the parent who spends most time with the child and knows the child best (primary parent). In most cases this will be the child's biological mother,

but may at times be the biological father, a stepparent, an adoptive parent, a guardian, or anybody else with a parental relationship to the child. The P1N is divided into 13 sections. These are:

A) Child's Family Details;

B) Health;

C) Education and Child Care (just 'Child Care' for infant cohort);

D) Relationship History;

E) Non-Resident Parent (only filled out where applicable);

F) Parenting Practices;

G) Sources of Support;

H) Parental Background;

J) Paid Work;

K) Financial;

L) Housing and Neighbourhood;

M) Direct Assessment of Child; and

N) Interviewer Observations.

While the questions are mainly answered by the primary parent, this is not always the case. Most questions in sections H, J, K and L may be answered by either one or both of the child's two in-home parents. Sections M and N also do not require responses from Parent 1.

Parent 1 Self-Complete Questionnaire (P1SCQ)

The P1SCQ contains questions that are best answered by written questionnaire rather than direct interview. The P1SCQ contains 5 sections. These are:

A) Child's Personality and Behaviour (just 'Child's Personality' for infant cohort);

B) Family and Community;

C) Combining Work and Family (only completed where applicable);

D) Health and Lifestyle; and

E) Couple Relationships (only filled out where applicable).

Parent 2 Self-Complete Questionnaire (P2SCQ)

The P2SQ is only filled out for those children whose primary parent has a partner living with him or her in the home (whether or not this is a biological parent of the study child). The P2SCQ contains some questions from the P1N and P1SCQ where information on both parents is required. It contains 5 sections. These are:

A) Parenting;

B) Combining Work and Family;

C) Health and Lifestyle;

D) Couple Relationships; and

E) Family and Relationship History.

Time Use Diary (TUD)

The Time Use Diary (TUD), completed by the primary carer, collects data on a child's activities throughout two 24-hour periods divided into 15-minute blocks. For each 15-minute block, options are presented in four categories. These are:

A) What the child was doing;

B) Where the child was;

C) Who was in the same room, or nearby if outside; and

D) Whether someone was being paid for this activity to take place.

An additional page of questions about the process of filling in the diary and whether the day was atypical for any reason is also included.

Family Contact Form (FCF)

The FCF records information about any contact between the interviewer and the family of each of the selected children, whether they go on to participate in the study or not. The information is mainly used by the fieldwork agency and the only information from the FCF which will be available in the publicly released dataset is information on the family's home and neighbourhood.

Home-Based Carer Questionnaire (HBCQ) (infants only)

The HBCQ is filled out only for those infants whose main type of regular non-parental care is used for more than 8 hours/week and is provided in a home environment. Parent consent is required to contact the carer. Home-based carers are

most commonly grandparents, but can also be other relatives, friends, neighbours, nannies, Family Day Carers, registered home-based carers, and other types of carers. The HBCQ seeks to determine the characteristics of the home-based care provided for the child and to gain further assessments of the child's characteristics from a source other than a parent. Its questions focus on the characteristics of the carer, the care environment, and the child's characteristics and behaviour while in care. It contains 6 sections. These are:

A) Care Arrangements

B) Background Information;

C) Child's Abilities and Behaviour;

D) Activities with Children;

E) Care Environment (only filled out for non-relative carers); and

F) Family Day Care Questions (only filled out for Family-Day Carers).

Centre-Based Carer Questionnaire (CBCQ) (infants only)

The CBCQ is only filled out for those infants who spend 8 or more hours/week in at least one type of regular care, and their main type of regular care is in some kind of child care centre. Parent consent is required to contact the carer. Centre-based carers work in long day care programs in centres, schools, occasional care programs, multi-purpose centres and other arrangements. Similar to the HBCQ, the CBCQ seeks to determine the characteristics of centre-based care and to gain further assessments of the child's characteristics from a source other than a parent. Its questions focus on the characteristics of the carer, the care environment, and the child's characteristics and behaviour while in care. It contains six sections. These are:

A) Service Characteristics;

B) Group Characteristics;

C) Personal and Professional Background Information;

D) Program and Environment;

E) Child and Family Focus; and

F) Child's Abilities and Behaviour.

Note – only one of the HBCQ or CBCQ can be completed for each infant, not both.

Teacher Questionnaire (TQ) (Four year olds only)

The TQ is answered for children in the 4-5 year cohort who attend a school, preschool, kinder or long day care centre. Parent consent to contact teachers is required. Similar to the HBCQ and CBCQ, the TQ seeks to determine the characteristics of the early educational programs a child is attending and to gain further assessments of the child's characteristics from a source other than a parent. Its questions focus on the characteristics of the teacher, the characteristics of the program, and the child's characteristics and behaviour while in the program. It contains 6 sections. These are:

A) Group Characteristics;

B) Child and Family Focus;

C) Child Skills and Competencies

D) Service Characteristics;

E) Background; and

F) Teaching Practices and Program.

NCAC Data Linkage

A key research question in LSAC relates to the impact of child care on children's developmental outcomes over time. While LSAC will collect parent-report information on children's child care histories and carer reports on the child care environment, relatively little systematic information will be collected on quality of child care due to a lack of resources within the data collection process.

The National Childcare Accreditation Council Inc. (NCAC) have quality assurance data on every Long Day Care (LDC) centre and on some Family Day Care (FDC) schemes. It is proposed that the LSAC dataset will include linked NCAC data for each child using LDC or FDC. This will ensure the LSAC dataset has good information on the quality of care that child is receiving. Without this empirical data on quality information, LSAC will have limited quantitative data to measure the possible impact of the quality of child care on future children's outcomes. Future analysis may link these data with other key variables as clusters of indicators for particular outcomes, such as school readiness, for these children. The potential of having such data is to provide good evidence to plan policy to achieve good outcomes in the transition to school and through the school years for all children.

Data to be collected from NCAC will include:

- Date of accreditation;
- Date of validation;
- Accreditation status;

- LDC and FDC schemes are assessed by staff, the centre director, parents, a validator and a moderator on 32-35 principles. These assessments will be available to restricted researchers.
- These assessments are weighted and then are used to construct quality ratings. Weightings will be available to restricted researchers.
- Length of accreditation;
- Demographic data on scheme;
- LDC data; and
- FDC data.

All the data items from the link with NCAC data will be in the unconfidentialised dataset and a small number of aggregated/composite derived items such as quality indices matching those used in the international child care research literature will be available in the moderately confidentialised dataset.

Variable Naming Conventions

Within the proposed data structure, the naming of variables within LSAC is critical, not only because of the vast number of items collected, but also because of the successive collection of data items from the same respondents over time.

Naming Standard

The naming standard needs to:

- be easily understood;
- provide as much information to the data user as possible; and
- maintain a maximum length of eight characters to accomodate all statistical analysis software programs.

It is proposed that the variable names follow the following format:

B n I S nnnx

Where:

B: refers to the cohort.

n: refers to the wave

I: refers to the survey instrument

S: refers to the section on the instrument

nnnx: refers to specific questions on the instrument. Generally nnn is the question number; and x is a sequential alphabetic indicator for the subpart of the question or response.

Variable Naming Conventions					
Position	1	2	3	4	5 -> 8
	B	*n*	*I*	*S*	*nnnx*
		Wave	*Survey Instrument*	*Section*	*Question Number*
	0yo Cohort: B 4yo Cohort: K Both Cohorts: D	Wave 1: 1 Wave 2: 2 Wave 3: 3…	P1F2F: C P1SC: P P2SC: S Teacher: T HB Carer: H CB Carer: L	Section A : A Section B : B Section C : C...	1 --> 999z
Key	*Cohort* B: Infant K: Four year old D: Demographic variables shared by both cohorts		*Survey Instrument* C: Parent 1 Interview – Child Data P: Parent 1 – Self Complete S: Parent 2 – Self Complete T: Teacher – Self Complete H: Home Based Carer – Self Complete L: Centre Based Carer – Self Complete		

For example, K3CA12 refers to those in the original four year old cohort at Wave three data collection. Data is from the Parent 1 Interview, Section A, question 12. Data associated with Dress Rehearsal x and Wave x are differentiated by person ID; that is, ID numbers for each study child differentiate between Dress Rehearsal participants and main wave participants. Dress rehearsal data will not be released. Data from the between Wave 1 and Wave 2 newsletter mailback will be released as part of the Wave 2 dataset.

Coding Framework & Data Dictionary

With the establishment of a clear variable naming standard, the process of organising and documenting each LSAC dataset includes the production of the coding framework and the 'Data Dictionary'.

Each data item will be clearly documented within the coding framework. The framework will list all data items that will be collected during Wave 1. Each survey instrument will have the following details specified:

i. the variable name (thus enabling users to cross-reference with the survey instruments, which will be available on the LSAC web-site);
ii. a name briefly describing each data item;
iii. a list of possible responses to each data item; and

iv. an indicator of the population covered by each item.

The information stored in the coding framework will be the base level of the LSAC 'Data Dictionary'. It is proposed that the data management team develop a meta-database with a web-based front end which could be accessed by all users, either by accessing the LSAC website or by loading the complete database to a local computer and using a web-browser to manipulate files. This product would be known as the LSAC 'Data Dictionary' and would comprise:

- Data items and their associated variable names (collected from the coding framework);
- Data items linked across successive waves;
- Data item descriptions;
- Identification of the construct being measured by data items;
- Thematic groupings of constructs;
- Thematic grouping across successive waves;
- Rationale for the implementation of each grouping – linked to the LSAC key research questions;
- The ability to search for items by question name/number, theme and variable name; and
- The ability to download material to assist in data analysis.

Such a product would circumvent the problems associated with multiple computer platforms and software choices made by the end user as most users have access to web-browsers. It would also add functionality in excess of most current longitudinal projects. The advantages of the product would multiply exponentially with each successive wave of data collection in terms of data item mapping, reference and manipulation.

Derived Variables

So far we have dealt with those core items that appear on the LSAC questionnaires. Among the data collected some items will be useful at this base level; others will need to be composited with other items to create new derived variables. Where this is a simple task (e.g. ensuring that weight, which may be collected in pounds or grams, is stored on the dataset only as kg/grams), the fieldwork agency will deliver a core set of derived items with the questionnaire data.

Complex Derivations

It is proposed that the AIFS data management team will create as many other, more complex, derived items as time and financial constraints will allow. These will be agreed to by FaCS in consultation with potential users. This will ensure consistency in approach by analysts. This will be an ongoing process which will increase in complexity with successive waves of data collection. Items to be added to the dataset will be specified by experts and consortia members. Many of these items have been tested during the Dress Rehearsal process. For example, physical health measures include:

- Weight for height percentile
- Change in weight for age z-score from birth to current

- Single waist measurement
- Birth head circumference z-score
- Birth head circumference percentile
- Birth length z-score
- Birth length percentile
- Birth weight for length z-score
- Birth weight for length percentile
- Length of post-birth hospital stay in days
- Current head circumference
- Current head circumference z-score
- Current head circumference percentile
- Average daily cigarette consumption of biological mother while pregnant
- Average days/week biological mother drank during pregnancy
- Average daily alcohol consumption by biological mother during pregnancy
- Stressful life events scale.

LSAC Outcome Index

Of the complex derivations to be output, priority will be given to the development of the LSAC outcome index. It is proposed that LSAC data be used to obtain developmental outcomes and to clearly reference benchmark points where children have poor outcomes at specific stages of development. The outcome index will be developed as a measurement of child development as reflected in the LSAC data. Drawing on the experience of the Canadian National Longitudinal Survey of Children and Youth (NLSCY), it will consider the concept of vulnerability. Children are considered vulnerable if they have at least one serious learning or behavioural problem (Government of Canada Applied Research Bulletin, 2001). The outcome index will be discussed further in future technical papers.

Other Longitudinal Studies

It is instructive to consider the approach to constructing more complex derived items taken by similar longitudinal studies. The extent to which other commensurate studies provide derived items is relatively consistent with the LSAC proposal. The UK Millennium Cohort Study includes a relatively small set of derived variables. Examples include gestation, birth weight, most recent weight, socio-economic classification and common household income. The Canadian National Longitudinal Survey of Children and Youth (NLSCY) also provides some derived variables of a similar nature.

Variable 'Keys'

In order to encourage uniformity and minimise computing efforts for the user, the LSAC Data Management team will also provide status variables which integrate data in a common variable showing the current status across a number of attributes for respondents. These files contain information about the survey data rather than the data itself. Its principal purpose is to support longitudinal analyses by allowing a user to restrict the dataset to the sample of interest, prior to matching of any cross-sectional files.

A critical attribute of Variable 'Keys' is that they assist researchers to define:

- the population of interest (e.g. basic demographics: family makeup, gender, cohort);
- the observation period; and
- the data structure.

Table 1: Some suggested variable 'Keys'

Variable name	**Meaning**
ID	Unique individual identifier (time invariant)
GENDER	Gender (longitudinally verified)
ORIGIN	Child's country of origin
FAMTYP1	Family type Wave 1
FAMTYP2	Family type Wave 2...
IDP11	Identification of Parent 1 Wave 1
IDP12	Identification of Parent 1 Wave 2...
IDP21	Identification of Parent 2 Wave 1
IDP22	Identification of Parent 2 Wave 2...
STAT1	Survey status Wave 1
STAT2	Survey status Wave 2...
CARE1	Child Care status Wave 1
CARE2	Child Care status Wave 2...

Demographic Data

The LSAC datasets will contain demographic information (employment status, parental and marriage/partnership biography, social background, etc.) gathered in the Parent 1 Face to Face questionnaire.

Attention will be given to the need for potential updating of demographic information, given some is time-independent (e.g. the year of first immigration to Australia, or occupation of father/mother when respondent was 4 years of age), in which case no update is necessary, or time-dependent with a potential need for updates (e.g. new marriages).

Study Child Unique Identifier

As previously mentioned, each study child will have a single, unique identification variable to ensure unique matching and merging across files and waves. In practice, each observation within LSAC, regardless of wave, will be able to be traced back to the original data of Wave 1.

Data Delivery

Although we have proposed that datasets be stored separately for each wave of data collection, information for several years can be merged together in a very straightforward manner for data delivery. Thus files can essentially be stored/administered as cross-sectional files, but distributed to the user as "ready to go" longitudinal files.

Recording of Missing Values

The assignment of specific values where instances of missing data are found is critical to the successful use of the LSAC dataset. Analysis is confounded where missing data is not correctly specified, estimates may be biased and quality checks of missing data are impossible without coding indicating whether the item was justifiably left empty. In the public-use version all missing data (whether it be because the respondent did not know or refused to answer, or because the question was not applicable and hence was not asked) will be coded into the following set of negative values:

-1 Not applicable (when explicitly available as an option in the questionnaire)
-2 Don't know (interviewer code)
-3 Refused or not answered (interviewer code)
-4 Value implausible (as determined after extensive checking)
-5 Unable to determine value
-9 Not asked: question skipped due to answer to a preceding question
. Missing data
* Invalid multiple response
# Scan Error – edit triggered

Ways of handling missing data are discussed in sections on imputation and weighting.

Confidentialisation & Access

The following outlines a draft strategy that will be further developed to guide data confidentialisation and access.

Classification of users

Due to the relative sensitivity of LSAC data, it has been decided to classify users of the data into various user groups based on the degree of risk associated with their access. This classification system will in turn dictate which datasets the user will have access to, and by which means (i.e. type of contract).

Classification of datasets

The LSAC datasets will be categorised as follows:

1. Moderately confidentialised – postcode data and date of birth removed and some top coding on variables such as income and number of children.
2. Unconfidentialised – postcode data and date of birth will be available and no data will be aggregated.

Researchers

Type I researchers

Academic researchers attached to a university, institute of technology or TAFE will be regarded as Type I researchers. This includes Consortium Advisory Group members, and honours or graduate students under the supervision of an academic. They could be granted access to both datasets and will be required to sign a license with FaCS to use the moderately confidentialised dataset or enter into a fixed term contract for access to the unconfidentialised dataset. FaCS will have responsibility for granting approval to the unconfidentialised dataset.

Type II researchers

Researchers attached to other organisations (such as private research organisations, lobby groups) will be regarded as Type II researchers. Such researchers seeking access to the data will be considered on a case-by-case basis. FaCS will have responsibility for granting this approval.

Overseas researchers

Overseas researchers will be permitted access to the moderately confidentialised dataset only. FaCS will have responsibility for granting this approval, which may include consulting with the FaCS Agency Security Adviser.

Government agencies

Commonwealth agencies

In recognition that all Commonwealth agencies (including Government departments and statutory bodies that operate under the Financial Management Act) are required to comply with Commonwealth privacy and security legislation and policy, greater flexibility will be employed in granting data access to such organisations.

State/Territory agencies

State/Territory agencies will be able to access both datasets under similar conditions as Type I researchers.

Statutory bodies

Statutory bodies whose primary business is research will be able to access both datasets under similar conditions as Type I researchers. Other statutory bodies will be granted access in the same way as Type II researchers; i.e. access considered on a case-by-case basis, with FaCS having responsibility for granting this approval.

Signatories to FaCS' social policy research agreements

FaCS currently has in place Social Policy Research Service agreements with three research providers.[1] These agreements specify the conditions under which FaCS unit record data will be released to these organisations. As such, access to the LSAC datasets may be granted to the relevant parties without the need for a separate contract.

Confidentialisation

We have a duty to respondents to ensure that the data they provide is not identifiable to those accessing the data. A large number of confidentialisation techniques can reduce the identifiability of an individual's data. However, with the implementation of each of these techniques, valuable data are lost. Achieving a balance between the protection of an individual's data and the production of a high quality dataset is always a complex balancing act and the LSAC team expects and welcomes input into fine-tuning the proposed approach. Before the implementation of any confidentialisation methods, the proposal will also need to be shown to meet the security standards required by FaCS.

De-identification

The first and most obvious means of protecting the confidentiality of any data received is for all name and address details to be removed. Upon receipt of name and address information from the Health Insurance Commission (HIC), I-View and NCS Pearson are bound by the same conditions as HIC under the Health Insurance Act; primarily, no personal information is to be passed on to any third party under any circumstances. The fieldwork agency will ensure that no data collected from any individual will be stored directly with contact information. The main datasets will contain identification numbers for each child, which can only be linked to specific personal details by the fieldwork agency. No released datasets will contain this personal information.

---

[1] The Melbourne Institute of Applied Economic and Social Research (Melbourne Institute) at The University of Melbourne; the Social Policy Evaluation, Analysis and Research (SPEAR) Centre at the Australian National University; and the Social Policy Research Centre (SPRC) at the University of New South Wales.

Even with the deletion of personal contact details, it is, however, difficult (if not impossible) to fully eliminate the possibility of self-identification by a respondent, or identification of another sample member when they are known to have participated (eg a spouse, a child or a neighbour). One method of protecting the confidentiality of de-identified datasets is to provide a restrictive access policy whilst retaining all data intact (i.e. un-manipulated, with no deletion of valuable data).

Thus, it would be possible to allow the final unconfidentialised dataset to be made available for specific use for Type I researchers and government agencies.
Before the data could be released, applicants would be required to apply for access (as outlined below) and to sign a contract with FaCS outlining the conditions under which data may be utilised. The purpose of the contract would be to ensure that, in relation to the datasets:

- the datasets are effectively utilised for FACS-approved research projects;
- appropriate data security arrangements are specified and adhered to;
- the parties each act in an effective and efficient manner;
- each party behaves reasonably and in good faith towards the other party; and
- the parties have a mechanism for managing use of the datasets and dealing with any disputes.

Parties granted a deed contract would be required to take all reasonable steps, and do all things that may be reasonably required by FaCS to keep the confidential information, including all documents, and all other things recording, containing, setting out or referred to any confidential information, under effective control of the licensee.

Moderately Confidentialised Dataset

For those researchers or institutions for whom it is not considered appropriate to access deidentified LSAC data, it is proposed that the dataset will be further confidentialised to reduce the risk that individual sample members can be identified. For the creation of anonymised datasets, various techniques could be applied - aggregation techniques, suppressing variables and ensuring there is a delay between collection and reference period of the data and its release as data.

Some possible methods include:

- aggregating geographic coverage;
- rounding;
- grouping or combining categories;
- top- (or bottom-) coding;
- imputing values from a model;
- suppressing fields or cells;
- suppressing variables; and
- time delay.

The implementation of any of these methods leads to a reduction in the data available for research, hence weakening the capacity to address research questions as

powerfully. An example would be the reduction of geographic specificity. To make data available only at the national level (a consideration when aggregating geographic coverage) would be to disallow research at the state/local level. Hence, minimising the use of these techniques is highly desirable, especially as datasets would only be distributed subject to the demands of the deed of license. In creating a confidentialised dataset, it is anticipated that the following processes will be involved:

i. some variables to be withheld (e.g., postcode);
ii. others to be aggregated (e.g., occupation is only provided at the two-digit level); and
iii. others to be top-coded (e.g., age and income variables).

A moderately confidentialised dataset subjected to these methods would then be available for research purposes to approved users through the same data access management regime; namely, through the application and granting of deeds of license.

Managing access to LSAC datasets

Data Managers

The Data Managers in AIFS and FaCS will be responsible for facilitating access to the datasets, monitoring adherence to security requirements, and contributing to a registry of users and completed research papers.

In the case of requests to access the moderately confidentialised datasets, the Data Manager will be a member of the LSAC (AIFS) team (with the exception of Type II and overseas researchers). For all other access requests, the relevant FaCS officer will be the Data Manager.

- FaCS will be responsible for signing the contracts and licenses governing the release of both the confidentialised and unconfidentialised datasets.
- The LSAC (AIFS) team will provide technical assistance to users of both datasets, as well as developing the meta data.

Requests to access the LSAC datasets

All requests to use the LSAC datasets must be made in writing to the Australian Institute of Family Studies (AIFS) using the appropriate form on the AIFS website. A request should include:

- the name of the organisation requesting the data;
- a description of the proposed project or general research purpose;
- the names and qualifications of researchers requiring access to the data under the project; and
- the project methodology and timeframe (necessary only if a specific project is proposed).

Access to datasets will be further subject to a determination, by the relevant Data Manager, that:

- the applicant is able to meet the control measures outlined in this paper; and
- the release is consistent with the purposes for which the data was collected and that there are no impediments to the data being used in this way.

It is acknowledged that researchers may not have a definitive project proposal in mind when applying to access the LSAC datasets. However, in order for FaCS to monitor and record the use of LSAC data, access requests should at minimum provide details of the general purpose of the research, as well as addressing the relevant points listed above.

Access to the datasets will incur an administrative fee of around $80 per CD and will be charged to each individual wishing to access the data. The fee will apply from April 2005 (Wave 1 public data release) and will be payable to the Australian Institute of Family Studies.

Users must notify their Data Manager if there is any change in their place of employment (including retirement) or in their research topic, as such changes may impact on their access status.

Managing the application process

For the moderately confidentialised dataset, the process will be as follows:

1. prospective users will send an application form and signed deed of confidentiality (downloaded from AIFS website) to AIFS.
2. The Australian Institute of Family Studies will assess suitability of user.
3. If user is approved, AIFS will send the license to FaCS. If user is an overseas or non-academic researcher, AIFS will forward the application form to FaCS.
4. FaCS will sign license and send a copy back to AIFS.
5. AIFS will be responsible for monitoring adherence to security requirements.

For the unconfidentialised dataset, FaCS will be responsible for processing the application, signing the contract and monitoring adherence to security requirements.

Registry of users

Data users will be required to provide the AIFS or FaCS Data Manager with details of any work utilising the LSAC datasets (including advance copies of any papers produced). AIFS will maintain a registry of users and their work (completed and in-progress). FaCS will be responsible for submitting to AIFS details relating to users of the 'unconfidentialised' dataset.

This registry and copies of research findings (including work-in-progress) may be published on the LSAC Internet site and in the LSAC annual report.

Clearance of research papers

Non-Commonwealth government researchers using the unconfidentialised LSAC datasets will be required to clear any papers for publication with FaCS to ensure the privacy of the survey respondents has been maintained. All researchers using the moderately confidentialised datasets will be asked to provide AIFS with an advance

copy of any research to be published. AIFS will pass a copy of the research to FaCS but clearance by FaCS will not be required.

User Support

With the datasets supplied to users, the following documentation will be provided in the form of a user manual which would include:

- Description of the conduct of the survey;
- Details of weighting and imputation procedures;
- Questionnaires;
- Variable listing outlining variable names, labels, response categories;
- Details of derived variables;
- Structure of the datasets; and
- Examples of use of the datasets.

User Training

In conjunction with the public release of the dataset after Wave 1, user training sessions will be offered by AIFS to further develop the information provided in the user manual and to allow users to dynamically interact with the LSAC Data Management team.

Data Security

Legislation

The LSAC project operations team and contractors I-View and NCS Pearson will administer personal information in accordance with legislation and policy including:

- Privacy Act (1988)
- Health Insurance Act (1973)
- Crimes Act (1914)
- Commonwealth Protective Security Manual (2000) [PSM]
- Australian Communications – Electronic Security Instruction 33 [ACSI]
- National Health Act (1953).

LSAC materials (including the datasets) are data owned by the Australian Government Department of Family and Community Services and as such will be handled in accordance with Commonwealth legislative requirements.

LSAC and the Information Privacy Principles

Eleven Information Privacy Principles (IPPs) are specified under the Privacy Act. These are concerned with how one collects, accesses, stores, protects, uses and discloses personal information. All processes dealing with personal data in the *LSAC* project will be subject to these principles. A summary of the IPPs is presented in Table 2. I-View and NCS Pearson are also bound by the National Privacy Principles where the IPPs have no equivalent provisions.

Table 2 Summary of the IPPs

	Principle	Action
1	Collection of personal information	Personal information shall only be collected where it is for a lawful purpose directly related to a function of the collector. Information should be collected from the individual where it is reasonable and practical to do so.
2	Informed participation	Where personal information is collected, participants will be made aware of the purpose of the collection and to whom information may be passed.
3	Data integrity	The organisation will take reasonable steps to ensure that personal information is accurate, complete and up to date.
4	Data security	Reasonable steps will be adopted to ensure against loss, unauthorised access, use, alteration or disclosure of personal information. When personal information is no longer needed it shall be destroyed or have all personal identification removed. (The National Privacy Principles also state that government assigned identification cannot be used to identify individuals).
5	Accountability	The Privacy Commissioner will be provided with an annual statement regarding the types of personal information held by the organisation, the purpose for keeping it the steps that should be taken by persons wishing to obtain access to that information.

		annual statement regarding the types of personal information held by the organisation, the purpose for keeping it and the steps that should be taken by persons wishing to obtain access to that information.
6	Access	Individuals shall be given access to their personal information after following strict protocols confirming the individual's identification.
7	Accuracy	Records shall be scrutinised to ensure accuracy. These records will be able to be amended by correction, deletion or addition.
8	Record usage	Prior to using a record containing personal information, data shall be checked for accuracy, currency and completeness.
9	Appropriateness of usage	Personal information will only be used for purposes for which the information is relevant. This information shall not be used for any other purpose than that for which it was collected.
10	Limits on use of personal information	Data shall not be used for any other purpose than that for which it was collected unless the individual concerned has consented or such use would lessen a serious and imminent threat to the life or health of the individual.
11	Limits on disclosure	Information shall not be disclosed to a person, body or agency other than the individual concerned unless the individual has consented, such disclosure would lessen a serious and imminent threat to the life or health of the individual, the data is required under law or is necessary for the enforcement of criminal law.

Within the framework of IPPs, the *LSAC* project team and contractors are obliged to:

- ensure that processes are in place to ensure any confidential/personal information is held in safe custody at all times;
- ensure that data are protected from all reasonably foreseeable loss, unauthorised access, use, modification, disclosure or other misuse;
- ensure that the identity of the respondents is not disclosed;
- ensure that all persons who have access to confidential information have executed a Deed of Confidentiality;
- cooperate with reasonable demands or inquiries made by the Privacy Commissioner;
- provide a policy on the management of personal information to requesting respondents.

Undertaking to Respondents

The explanatory brochure distributed to all participants of the survey gave the following undertaking with respect to their privacy and how the information would be used:

> *It is natural for you to be concerned about how your privacy will be protected when you are involved in the study. Some ways we will protect your privacy are:*
>
> - *Before any information is used, all of the details which could identify you or your child will be removed.*
> - *Only group data will be reported, and we will not identify individual cases.*
> - *Very strict procedures are in place to make sure only authorised persons have access to the data. All interviewers and researchers working on the data are required to sign a Deed of Confidentiality.*
> - *The study has been approved by the Australian Institute of Family Studies Ethics Committee, and all interviewers and researchers involved must comply with the Privacy Act 1988.*

Further information on LSAC security obligations can be obtained by contacting the project operations team at AIFS.

Data Imputation

Imputation refers to the replacement of missing data with a substitute that allows data analysis to be conducted without being misleading. The major application possible for LSAC is the estimation of attributes where the interviewee refuses to answer the survey question or the attribute is missing. It is also possible to estimate the attributes of children not available for interview.

The following have been identified as key components of successful data imputation by the HILDA data management team. These will be adopted as the driving principles of the LSAC imputation policy:

i. Imputation should not lead to biases or distributional changes in the data, or significant extra variance to estimators.
ii. The imputation process should rely on data from the sample rather than making external assumptions about the likely nature of missing data.
iii. Imputation should not lead to important sample estimates being based too heavily upon imputed values.

Non-response

Evidence from the LSAC Dress Rehearsal suggests that item non-response was minimal for responding units but that unit non-response was significant. The size of the study requires that a strategy be established to effectively deal with missing data and non-responding units.

The treatment of non-response requires an understanding of the characteristics of the non-respondents and the likely impact these non-respondents would have on the survey data. It is proposed that LSAC treats unit non-response through weighting and item non-response through imputation. At times it may also be appropriate to ignore non-respondents – effectively assuming that non-respondents are like respondents or that any bias introduced by excluding the non-respondents from the analysis will be very small.

Imputation in other Longitudinal Studies

The use of imputation in similar studies overseas has been minimal. The National Longitudinal Survey of Children and Youth performed limited imputation to their datasets, restricting it to missing values in variables used to calculate scales. The Millennium Cohort Study has not implemented any imputation within delivered datasets, preferring researchers to access an unmanipulated but 'clean' dataset.

HILDA's investigation of imputation in other Australian longitudinal studies has also shown minimal useage to date. The General Customer Survey and Longitudinal Data Set, both run by FaCS, do not as yet include any imputation. For the Longitudinal Survey of Australian Youth, the treatment of missingness was left to researchers. In the Longitudinal Study of Immigrants to Australia, item imputation was not done and unit non-response was dealt with through weighting. The Longitudinal Study of Women's Health does not routinely impute missing data. The Survey of Employment

and Unemployment Patterns, undertaken by the ABS in 1994 to 1997, used imputation for wave non-response, but did not use imputation for item non-response. Information from earlier waves was used to construct imputation classes and a 'donor' from the same wave was identified using the 'hot deck' procedure.

LSAC – Whether to Impute

Based on other national and international longitudinal studies and the difficulties involved in longitudinal data imputation it would appear that either (a) no imputation should be conducted on the LSAC dataset and that researchers make allowance for missing data in their analyses; or (b) a minimalist imputation strategy should be adopted, primarily based on items exhibiting missingness levels of less than 10%.

In determining which option to adopt, the Data Management team will rely on input from FaCS, consortium members and future users of the data.

The following discussion focuses on possible strategies if option (b) is adopted.

Principles

It is proposed that the guiding principles for any LSAC data imputation strategy should encompass the following:

- Only items at the specified threshold of missingness (5-10%) will be considered for imputation.
- Any imputed variables should be clearly identified such that the users can use the imputed variable or original variable as they wish.
- The imputation should maintain, as far as possible, the underlying variability in the data.
- Imputation be used for scale construction where items are missing.

The LSAC imputation strategy will target specific variables with levels of missing data between five and ten percent. This proposed strategy is based on the following:

- Where there is a small amount of missing data the existence of item non-response is likely to be more nuisance value than substantially affecting analyses.
- By imputing these missing values, results of analyses are unlikely to change but it will ensure data is more 'user friendly' in that the 'missing' line can be eliminated from tables, thus avoiding any complication in table percentage calculations.
- Furthermore, the need to drop records when running regression models is avoided.

In essence, no harm can be done by imputing for small amounts of non-response but one is able to produce a more useable dataset.

However, where there is a significant amount of non-response (>10%) there is a potential for non-response bias beyond what can be corrected for in an imputation strategy. Imputation cannot create data where it doesn't exist. It doesn't increase the sample size, although imputing for a large amount of missing data can give analysts a

false impression of the sample size for an analysis. For any data item exhibiting missingness at such a level (i.e. >10%) it is proposed that imputation should not occur. Rather, end users of the dataset will be provided with reference information about such items. This information will seek to explain the low response rate and may investigate areas such as:

- whether the item was one that a particular group of respondents was unable to answer;
- whether the item was confusing;
- whether the information requested was too sensitive.

Candidates for Imputation

It is proposed that for those data items collected in Wave 1 that show a level of missingness between five and ten percent, that discussion will be entered into with management at FaCS to determine a priority listing of imputation candidates. It is acknowledged that LSAC is limited by the amount of resources that can be spent on imputation. Therefore, we propose to restrict our attention to a subset of variables exhibiting missingness levels between five and ten percent with priority given to those missing valuables used in the calculation of scales.

Imputation over Time

As LSAC is a longitudinal study, the method of imputation needs to accommodate the impact of imputation over time. Many imputation techniques have been developed for cross-sectional surveys and the emphasis has been on population estimates and variances at a point in time.

When repeated observations are collected over time, the estimates of change between waves are very important. It is important to avoid any introduction of variability into estimates of change by imputing solely based on cross-sectional information, thereby suggesting there has been change when there has not been.

Furthermore, decreasing the variability of the change estimates by imputing solely from information about an individual (such as carrying forward the last observed value), which would suggest there has been no change when there may have been some, and thus should also be avoided.

It is possible that longitudinal information may be used in the imputation process. This would involve a recalculation of imputed values at every wave. However, whilst the use of longitudinal data may improve the ability of a predictive model to impute data, there would be no single, master dataset for any one wave. Longitudinal imputation options will be assessed in later LSAC technical papers.

Imputation Method

Adapting the experience of the HILDA survey for the purposes of LSAC, and acknowledging that the majority of variables within LSAC are categorical, the proposed imputation method is the ‘hot deck’ procedure. This technique appears to be

most appropriate due to its ease of use and ability to maintain the variability in the data.

The ‘hot deck’ method divides the complete cases into imputation classes based on key variables (e.g. family type). The incomplete case (i.e. the recipient) can then be matched to an imputation class from which a case with complete information (i.e. the donor) can be chosen. The missing data is then replaced with the valid data from the donor case.

If a significant number of continuous variables are found to require imputation, it is proposed that, for these variables, it may be more appropriate to implement the ‘nearest neighbour’ method of imputation. This method is an extension of the hot deck procedure where the distance between the non-respondent and respondents is calculated based on the observed variables and the closest respondent is chosen as the donor for the non-respondent.

Both the imputed and the original variables will be provided on the datasets. The imputed variable will contain the original data for the non-missing cases and the imputed data for the missing cases. Researchers will then be able to choose which variable they wish to use. It is felt that this approach is better than having one variable with an imputation flag as researchers do not have to do any work in creating the original variable and the reason for the non-response is not lost.

Outliers

In dealing with outliers in a dataset, two problems present themselves. The first is identifying outliers and the second is deciding what to do with them.

Inspection of frequencies, and at times histograms for continuous data items allows identification of univariate outliers. It would also be useful to check for multivariate outliers when the variables have some obvious common variance (eg employment status and income) where the presence of outliers is likely to point out obvious errors. In most cases, this will only involve the interaction of two variables, hence scatterplots, split file frequencies or histograms, should identify them. However occasionally more complex checks may require calculation of a Mahalanobis distance through regression analyses. Mahalanobis distance is used to identify cases that have an unusual pattern of scores.

What to do with outliers once identified is more controversial. Tabachnick and Fidell (1996) identify four reasons for outliers in a dataset:

1. Incorrect data;
2. Unspecified missing data;
3. Data which does not come from the target population intended to be sampled (i.e. the person should not have been asked the question);
4. Correct but extreme responses, which may add unacceptable variability to estimates.

Most of the outliers found in the LSAC datasets are likely to be due to the first reason, and hence all data identified in outlier searches will need to be thoroughly re-checked. If a data item is clearly incorrect then the rawest form of the data will be checked. It is proposed that if there is no indication as to the true value it must be considered missing.

The second category can be easily discovered and corrected, while the third should be identified by our skip checks and deleted accordingly.

The fourth reason is clearly the most complex since the data identified accurately reflects the "real world" status of the variable it describes (eg some people do earn $1,000,000+ a year when most people's annual income is measured in the tens of thousands). Tabachnick and Fidell (1996) recommend two potential courses of action here: a) altering the data to make it less extreme, b) transforming the variable to make the extremity less important.

For the purposes of LSAC, however, neither of these options is desirable since it cannot be anticipated how the data will be used once it is made public and an optimal solution for one analysis might not be best for another - it is rare in a survey that just one estimate is produced. Hence, clear criteria for unacceptable variability cannot be easily determined. However, a common principle is that no single data value should contribute more than a certain proportion of specified key estimates. For example, it might be decided that no single household income value should contribute more than

10% of a state estimate of average household income. It is unlikely that extreme values will contribute more than 10% of estimates because weights will be relatively uniform. It should also be noted that the majority of LSAC variables are categorical. Nevertheless, in the detection of outliers found to contribute significantly to estimates (i.e. >10%), confidentialisation processes will be followed.

Once the decision is made that an extreme value is unacceptable by such a criterion, it is usual practice to replace it with the most extreme acceptable value. Typically this introduces a bias through the removal of the 'excess contribution'. It is possible but not common practice to then reallocate this contribution to other sample units to maintain unbiasedness of estimates.

It is proposed that within the LSAC study, outlier treatment should use this principle of the maximal allowable contribution of any single data record to key estimates. The criteria should be developed through an analysis of Wave 1 data. It is anticipated that very few items will contribute to greater than 10% of estimates.

Weighting of Data

LSAC aims to collect data from a representative sample of Australian infants aged up to 13-14 months and children aged 4-5 years. This sample is drawn from those children registered on the Health Insurance Commission's Medicare database. This database has the most extensive coverage of children in the desired age range of any Australian database. However no survey can ensure that the collection process is perfect and this is especially the case with a longitudinal survey where, even if the original sample is accurately representative of the population, bias may occur across successive waves. Hence there is a need to adjust the data so that it continues to accurately reflect the population of persons being surveyed. If a sample was ideal, all the weights might be equal to the ratio of the population size to the sample size. However in reality they will be unequal to compensate for sample design constraints, operational difficulties and, most importantly with a longitudinal survey, the way that the sample alters over time.

The main purpose of weights in this study will be to compensate for differences between the population on the sampling frame and the 'real' population, as best estimated from ABS data, differential initial non-response and, in subsequent years, differential attrition. It is important to remember that weighting can only be based on characteristics for which the distribution across strata in the population is known.

To allow for any non-response effects, a post-stratification weighting system using population benchmarks derived from ABS population estimates will be investigated, in conjunction with the Health Insurance Commission non-response information. Separate weights could be determined for each of the cohorts within each regional stratum, based on the appropriate sub-populations and response rates. Within the limitation of what a single set of weights can do, the weights should attempt to give proper weighting to each child in each wave, so that each wave can be used as a valid and representative cross-sectional dataset.

Principles

In considering weighting it is worth defining the principles against which procedures can be judged. The following have been identified by the HILDA data management team:

i. The weights should be considered as expansion factors permitting the scaling of the sample to the population. Hence the sum of the weights should accurately match known population parameters such as the total population.
ii. The weights should adjust for unequal probabilities of inclusion in the survey, to redress any potential sampling biases. In many cases this will mean that weights are inversely proportional to the probability of inclusion.
iii. Ideally weights should not vary from a constant value more than can be avoided since this reduces the statistical efficiency. This is obvious if the situation of spending significant effort on collecting data for a particular unit and then giving it very low weight – much of that effort is effectively wasted.
iv. Where certain analyses may be restricted to subsets of the population, specialised weights may be required. However such weights should be as

consistent as possible with the principal weights since they should not lead to contradictory analyses.

Since similar issues relate to LSAC, these principles will be adopted.

Issues Affecting Weights

Initial sampling

The first factor to impact upon weights is the initial sample design that determines which children are in the study in Wave 1. Hence the initial *design weights* may be uniform. However it is likely that variations in design weights may be required to account for:

i. the actual number of children enumerated per Postal Area being different from the number used in the process of selecting Postal Areas; and
ii. response rates varying from those expected.

When the sample is established in Wave 1 it will be a (close to) representative sample of children in the two age ranges. Over time changes will take place as members leave the sample. It is not automatic that these changes allow the sample to remain representative. The importance of weights in a longitudinal survey is that they are the only means of adjusting an evolving sample for the drift of the distribution away from that of the population.

Types of Weights

Weights will always depend upon the type of analysis being carried out, but there will always be two basic types depending upon whether they relate to children and cross-sectional or longitudinal studies.

Response rates

At various stages in the survey, parents may refuse to participate. If the unobserved characteristics of the refusing individuals are the same as those who participate, this is not a significant problem. In practice however there will be relationships between the child/parent's characteristics and the probability of refusal. Hence the weights should attempt to adjust for any such biases.

This is typically done by dividing the sample into response classes and weighting to ensure that each response class has an appropriate final weight. For example, if non-response leads to single mother households being under represented in the sample, then those that are in the sample should be given higher weight.

Attrition

Like initial non-response, attrition is likely to be related to the characteristics of the children being studied. It is likely that factors that lead to attrition – family disruption, employment changes and relocations – are variables of significant interest in the survey creating a situation where persons of greatest interest may well be the most difficult to collect longitudinal data for.

Weighting Approaches in Similar Studies

Within the Millennium Cohort Study the sample of births selected was clustered geographically and disproportionately stratified to represent areas with high proportions of ethnic minorities in England, residents of areas of high child poverty and residents of the three smaller countries of the UK (Wales, Scotland and Northern Ireland). The sample is thus a disproportionately stratified cluster sample. The disproportionality means that the sample is not self-weighting and so weighted estimates of means, variances etc. are needed. The weighting that was performed on the sample was done to compensate for the deliberate oversampling. This weighting was based on the probability of a child being selected for the study, not the characteristics of the final sample.

Within the Canadian National Longitudinal Survey of Children and Youth, weighting has been carried out to give expected population frequencies for all items; that is, estimates are generated that are predictive of the results for the population.

Calculation Procedures

For a longitudinal survey two sets of weights can be considered: cross-sectional weights and longitudinal weights. Each wave can be considered as a cross-sectional survey in its own right. After the first wave, cross-sectional analyses will be the only analyses that are possible, and cross-sectional weights will be important for the first wave. As future waves are conducted, the analytic focus may shift towards the longitudinal analyses of changes and trends, but there will always be the capacity to produce point estimates from individual wave datasets.

For cross-sectional weights, the goal is to accurately estimate population quantities. The weighting needs to take into account three factors:

i) differences between the sample frame and the target population of interest;
ii) probabilities of selection of individual children;
iii) the impact of non-response.

The target population of interest for each wave may be defined as the population of children within the specified age ranges living in Australia. The sample frame, being the HIC data base, may differ from this desired scope by virtue of missing registrations. The differences in scope can be accounted for by benchmarking the weights to estimates of the total number of children in each target population. The ABS *Estimated Resident Population* series may be used for this purpose.

Within each cohort, the LSAC sample has been selected with approximately equal probability of selection for each in-scope child in each stratum. The stratum sample sizes have also been chosen to keep probabilities of selection approximately constant across strata. Postal areas have been selected with probabilities proportional to the number of in-scope children, and then a fixed cluster size has been selected within each postal area. This *self-weighting* design would yield equal weights for each child if there were 100 per cent response.

The experience of the pilot test and Dress Rehearsal has shown that non-response will be a significant issue in the survey. The most important part of the weighting strategy will be to account for this non-response and biases that could be introduced because the non-respondents are not a random sub-group of the selected sample. Before finalising the weighting strategy, it will be important to analyse the non-response to the first wave, to try and detect any patterns that might suggest a non-response bias. This can be undertaken by:

i) analysing response rates and non-respondents in comparison to HIC benchmarks as provided in the HIC statistical abstracts.
ii) collecting information about reasons for non-response, and analysing any geographical trends.
iii) comparing the demographic distribution of each cohort sample to the distribution of the population from the most recent census of population and housing. Variables that have been found to be associated with survey non-response in other household surveys include: age and employment status of the childrens' carers, family structure, socio-economic status, English language proficiency, and family size. These variables and other potentially relevant demographic factors should be examined.
iv) at the postal area level, response rates can be examined in comparison to the demographic composition of each area as measured by factors such as the Socio-Economic Indexes for Areas (SEIFA).

Depending on the results of this analysis there are three possible approaches to the weighting strategy:

i) if the non-response is judged to be random, then the design weights can simply be increased by a constant expansion factor to adjust for the proportion of non-response.
ii) if a small number of factors is found to be linked to participation in the study, the responding children can be post-stratified by these factors, and weights set within each post-stratum.
iii) if the number of factors linked to survey participation means that the post-strata would be too small to be able to calculate reasonable, consistent weights not subject to random fluctuations, then the calibration approach of Deville and Sarndal (1992) can be employed. In this approach, the design weight is taken as the starting point, and calibrated to give correct population totals for each factor included in the weighting, while minimising the deviation in weights from the original design weights. This strategy has been successfully employed in a number of surveys conducted by major statistical organisations such as the ABS and Statistics Canada.

Weights after Wave 1 – A modelling approach

At each wave the weights attached to individuals should inversely reflect the probability that that individual had of being in the sample. This would be the child's

Wave 1 weight appropriately modified by a factor inverse to their probability of *not* dropping out through attrition.

At each subsequent wave it is possible to calculate cross-sectional weights for that wave. That is, each subsequent wave can be considered to be a separate cross-sectional survey in its own right. The weights can be calculated using the same process as for the Wave 1 weights, with two additional modifications:

i) the pattern of response can also be analysed in terms of the distribution of responses for items collected in previous waves. This may yield further insights into the mechanisms underpinning the non-response that could be incorporated into the weighting strategy.
ii) if calibration is used for determining weights, the cross-sectional weight from the previous wave can be used as a starting point rather than the original design weight.

When calculating cross-sectional weights for the subsequent waves there are two additional issues to consider:

i) should population benchmarks be adjusted for migration in and out of the sampling frame? Children who were not living in Australia at the time of the sample selection, but who moved to Australia and are in the age range could possibly be represented in population benchmarks. An assessment would have to be made as to whether children who fall into this category could be assumed to be similar to other children selected in the survey, or whether they are systematically different.
ii) Children and families who migrate interstate, for example, and thus move out of the stratum they were originally selected in may need to be reweighted in regard to the stratum they are now residing in.

In addition to calculating cross-sectional weights for each wave, a set of longitudinal weights may need to be calculated. Rather than trying to represent an ever changing population at each point in time, longitudinal weights can take a cohort approach. The longitudinal weights would be based on the cross-sectional weights for Wave 1, but would be adjusted for attrition over time.

Role of survey weights in longitudinal data analysis

Surveys can be used to estimate characteristics of the population from which they were selected. Weights as described above are appropriate for calculating estimates of population prevalence for any characteristic of interest. They can also be used for descriptive analysis of demographic patterns in the characteristics of interest, and for describing differences between groups of children in cross-tabular analysis.

Another use of the LSAC data is modelling. Models can be fit to examine the associations between a range of items and to describe trajectories and trends over time. As the goal of modelling is to explore relationships in the population rather than to estimate prevalences in the population, the role of weighting is different. It may not be necessary to incorporate weights into a model. A sample design is defined to

be *informative* in respect of a given model if the model that would be fit to the full population data (were it possible to fit such a model) is different from the model that is fit to the sample data. A sample design would be informative for a particular model if there were a relationship between the outcome variable being modelled and the probabilities of selection *after* accounting for the explanatory variables in the model. If the factors that describe the pattern of non-response are not related to a given outcome variable, then the sample design would not be informative for that variable and weighting would be unnecessary. If say, the probability of non-response was found to be related to socio-economic status, and socio-economic status is included as a factor in a model, then it would not be necessary to make any additional weighting adjustment in the model.

Where the sample design is found to be informative for a given model, weights can be incorporated in the analysis using, for instance, the method suggested by Skinner and Holmes (2003). This involves using two separate weights: the original cross-sectional weight for the first wave, and an attrition weight calculated separately at each wave that represents the probability of remaining in the sample for that wave.

Recommendations

A consultant will be employed to assess the proposed weighting strategy. The following issues will be considered:

- The principal weighting scheme for LSAC should have weights that are comprised of the initial Wave 1 sample weight with adjustments applied at each Wave to adjust the longitudinal sample to population demographics.

- Population demographics may be applied to adjust weights.

- This weighting strategy will yield a dataset that accurately represents the two child cohorts. The adjustments will largely compensate for longitudinal evolution.

- Calculation procedures for weighting, imputation and outlier treatment should be implemented programatically to ensure correctness and objective application of procedures.

Respondent Tracking

As elaborated in the *LSAC Discussion Paper No.2*, the intent is that all children included in the Dress Rehearsal and first wave will be followed in all subsequent waves. The only exceptions are likely to be children whose families move overseas. Wherever possible, the primary parent will continue to be the principal informant on the child and family and it is likely that the other parent will also continue to be asked to provide data. Contact information will be sought from both parents independently during Wave 1 to assist in cases where there is a later separation of the parents.

The potential for non-response is present at every wave of a longitudinal study. Since there is detailed information on the characteristics of all respondents at Wave 1, it will be relatively easy to apply weights to the data to compensate for any bias resulting from subsequent non-response. However, such procedures are only likely to be effective in the short-run.

Over the longer term it is important to minimise attrition because of the probability that those lost from the study are different from the 'stayers' in ways that may not be observable at Wave 1. Furthermore, high rates of attrition have obvious detrimental effects on the sample size then available for longitudinal analyses of developmental trajectories and pathways to outcomes. Finally, there are good reasons to be concerned about the adverse effects of high attrition on the perceived legitimacy of continuing the study.

The experience of several overseas longitudinal studies indicates that attrition is likely to be highest in the early years. In the Canadian National Longitudinal Study of Children and Youth, for example, attrition between the first and second wave was approximately 11 per cent (NLSCY, 1999), but it is reported that retention has been maintained at 85-90 per cent over the later years (A. Zeesman, the Department of Family and Community Services Workshop, May 2001, Canberra). Similarly, in the Christchurch Health and Development Study the attrition rate was almost 9 per cent between birth and age 2, but dropped to less than 1 per cent per year subsequently, with a total attrition of 19 per cent by age 18 years (Fergusson et al 1989; Horwood & Fergusson, 1999). For these reasons, a minimum level of 85 per cent retention from wave to wave is expected across the entire life of the project. To achieve this will require implementing strategies that maximise the retention of sample within each cohort over the entire life of the project. The most successful sample retention strategies that are typically used are:

- inclusion of tracking questions in study instruments;
- maintenance and frequent updating of a database on respondents' location;
- promoting participant identification with the study; and
- extensive communication with sample members, including training interviewers in interviewee friendly techniques.

Information will be obtained from both parents on names, addresses and telephone numbers of 2 relatives or friends not living at the same address, as well as their own email addresses, and business and mobile telephone numbers.

Season's Greetings cards will be sent to all families, and birthday cards to all children annually, together with change-of-address cards for notification of any intended moves. Participating children will be given a small gift (bowl and cup) with the *Growing Up in Australia* logo, and attractive 'fridge magnets' will be left with parents with the study's contact details. Contact will also be maintained with participants between study waves through regular newsletters.

In addition, the study will be promoted through marketing of the logo and tagline, and through media exposure to the study, and a 1800 telephone number and website will be maintained so that participants can contact the data collection agency and AIFS.

If a family cannot be located through the contacts they have given, then forwarding addresses or telephone numbers will be sought from residents at the address or telephone number of the original sample member. If these means prove unsuccessful, the Electronic White Pages, Australia Post and the electoral roll will be accessed to pursue contact details for persons who have changed address.

In addition, the use of a brief between-waves mail-back survey in 2005 may help in maintaining contact.

Data Linkage

For a study such as LSAC, data linkage offers important potential to value-add to the primary modes of data collection from parents, carers, teachers and children themselves. It can:

- increase the efficiency of data collection, by avoiding the necessity to seek information directly from informants;
- reduce respondent and interviewer burden (for the same reason as above);
- provide information to which individual informants may not be privy; and
- access data at a higher level of aggregation than individual informants (e.g. at the community level).

Possible sources of data linkage are:

- The Australian Bureau of Statistics (ABS) where links can be made at the Census Collector District (CD) level thus enabling the inclusion of such variables as the community's socio-economic status, diversity/homogeneity, safety, housing, and the availability of relevant services and resources in data analysis. Such linkage does not require permission from parents. The use of the Global Position Sytems (GPS) in the LSAC study will assist in effective linkage at CD level data.
- National Childcare Accreditation Council (NCAC). The NCAC dataset contains detailed information on quality of care on all Long Day Care centres and Family Day Care schemes.
- Medicare records. Medicare activity data could give an indication of history of usage of some types of health services (held by HIC).
- Australian Childhood Immunisation Register (held by HIC).

Data linking will occur at a higher level of aggregation than individual informants in the first two sources. The latter sources would require matching at the individual informant level. The consent process for linkage to the latter sources requires permission from the child's parent or guardian. Linking with medicare records and the Australian Childhood Immunisation Register is not anticipated to occur at this stage due to lack of funds.

Data to be linked will be processed at I-View's offices and provided as part of the transfer of clean, deidentified data files. ABS type data will be linked by Census Collection District and data will be interrogated to ensure confidentiality is maintained. Data from these sources is currently available to I-View staff. The remaining data linking involves the matching of individual data. This will incorporate secure transmittal of subject consent forms and unique IDs to linked authorities whereupon data will be extracted and recorded onto CD ROM. This CD ROM will then be delivered by FaCS approved courier or PKI to the I-View offices where linking will occur in a secure environment. Any release of linked data will follow processes specified to ensure that anonymity of participants is preserved.

Software Implementation

Software Selection - A Comparative Overview

The selection of appropriate software is fundamental to a project's eventual effectiveness and efficiency. Consultation with local and international projects concerned with longitudinal data management has been critical in establishing the strategic direction of the current project. The following tables provide information and advice on database management and analysis.

Christchurch Health and Development Study

Data is provided in ASCII format, as it is readable by all statistical packages All editing and data manipulation is done in SAS. Researchers indicated that both SPSS and STATA provide similar capabilities Researchers recommended against Excel or Paradox for data management Researchers recommended having SAS, SPSS, STATA, a structural equation modelling program and a latent class package available for specific analyses Researchers recommended STATA's system of transferring files between different statistical programs; namely Stat/Transfer

Centre for Longitudinal Studies (National Child Development Study, 1970 British Cohort Study & the Millennium Cohort Study)

Researchers use Statistical Information Retrieval(SIR), SPSS, Excel and Access for data management. No specific program is used for analysis. Individual researchers choose their software (usually SAS, SPSS or STATA) Researchers are using SIR for data management of the National Child Development Study and SPSS for both the British Cohort Study and the Millennium Cohort Study. Researchers are currently looking to change to SIR for BCS70, as it has advantages for complex datasets STATA provides effective means of transferring files between statistical software packages via Stat/Transfer

Avon Longitudinal Study of Parents and Children

Minimal information forwarded. Researchers use SPSS for data management and both SPSS and STATA for analysis STATA provides effective means of transferring files between statistical software packages

Simmons Longitudinal Study

Researchers indicated that SAS and SPSS are the two market-leaders in the US and both are compatible with each other and largely equivalent Researchers stated that SAS is better for handling complex longitudinal research methodologies.

In their research, they use SPSS for data management and perform analyses in either SAS or SPSS

Minnesota Longitudinal Study of Parents and Children

Researchers use SPSS for data management and analysis
Additional analyses packages available include LISREL and MPLUS

US Child Development Supplement to the Longitudinal Study of Income Dynamics

Researchers use ORACLE for all data storage

US NICHD Study of Early Child Care and Youth Development

Researchers use SAS for management and analysis and consider it to have more flexibility than SPSS, but less user-friendliness

The Household, Income and Labour Dynamics in Australia Survey (HILDA)

The HILDA research team decided against implementing SIR as it has a small user-base, a slow development cycle and the long-term future of the production company was in doubt. Furthermore, there were concerns voiced by those running equivalent German and British longitudinal studies in SIR about their ability to replace key personnel
The researchers found that data being output from Surveycraft could separate 'not asked' questions in SPSS but not in SAS, so initial datafiles in SPSS format were superior
Researchers found that SPSS and SAS were functionally equivalent apart from this issue
Researchers decided that a codebook would be held in a metadatabase (produced initially in Microsoft Word). By Wave 2 release this metadatabase will be fully automated. The team decided to change the variable naming structure to deal with the fact that question categories were being changed over time, which required the renaming of about 1000 Wave 1 variables
For transfer of files into various formats (ie of interim datasets for those who need STATA or SAS) the researchers use Stat/Transfer v7

The Australian Temperament Project

Researchers receive the data in SPSS format and use the same program for data cleaning, derivation and analysis
Metadata is stored in a Word document
Variable names use the year of collection to differentiate collection waves
Client contact details are maintained in filemaker-pro

Women's Health Australia

Researchers receive the data as text and read it into SAS
Statisticians associated with the project generally use the SAS datasets but SPSS datasets are also created for researchers when required.
Metadata is stored in Microsoft Access, which was not a popular decision. If researchers do not have Access skills then analysis of data is problematic (i.e. the data dictionary is sent to collaborators with datasets, so if they don't use Access, variable interpretation is difficult)
The data dictionary now has about 2000 entries so itmust be in an indexed database
Researchers recommend SAS and not SPSS for data management
Researchers don't think SPSS has the same ease of data manipulation as SAS which is particularly important for complicated code
SAS data is easily outputed in SPSS format for other users.
Code is easier to write in SAS than SPSS

Dunedin Multidisciplinary Health & Development Study

Researchers conduct data management in SPSS, but STATA is used for analysis due to (a) its greater speed and (b) less superfluous output.
Researchers use a software product called DBMS copy to convert from one database format to another
Comment was made that many people prefer SAS but not enough was known about the product for the researchers to offer advice

Data Management

The comments suggest there are several adequate platforms among popular data management software choices. There appears to be some satisfaction with the most well known software but those involved in more complex studies are less encouraging of the adoption of SPSS as a tool for managing longitudinal data. Some researchers involved in more complex studies have reported that SPSS does not have the same ease of data manipulation as programs such as STATA and SAS. This is particularly important when one is dealing with complicated code inherent in longitudinal data management. Others involved in complex data management, such as the HILDA Management Team, have moved from SAS to SPSS mainly due to greater compatibility with 'Surveycraft', the software employed in data processing by HILDA's data collection agency.

The use of the less well known package SIR, whilst advocated by its users, would seem ill-advised as it has a small user-base, a slow development cycle and the long-term future of the production company remains in doubt. Furthermore, there have been concerns about the availability of personnel who are trained in SIR.

Overall, for the purposes of data management, STATA and SAS would appear the most viable alternatives for the current project. As knowledge of STATA processes is limited within the LSAC project operations team, common sense would suggest that the best known of two comparable packages be adopted; in this case, SAS.

Metadata

A problematic area in comparable studies' data management procedures has surrounded generating and maintaining comprehensive information on metadata. As the complexity of research projects increase, the ability to output complete and intuitive metadata guides is critical in ensuring ease of access for end users.

The size of the current study means that metadata management is a core consideration when assessing the merits of any overall software configuration. Previous studies have used *Microsoft Access, Microsoft Excel* and even *Microsoft Word* to manage metadata. The problems experienced in these software choices relate to end usability (e.g. If final users of the data are unable to operate *Access*, then meaningful analysis of the data becomes questionable). The use of *Microsoft Word* becomes less satisfactory as longitudinal studies progress, adding layers of complexity with each successive wave of data collection.

A metadata management program currently available to AIFS is FileMaker Pro. This program is similar to Microsoft Access in its ability to manage metadata. A metadatabase in FileMaker Pro would be capable of generating a codebook for each wave at each release. It would be also be capable of checking to ensure that all the variables in the datasets are in the codeframes and vice-versa. By publishing a codebook in a standard format (i.e. pdf), end-users avoid the need to be proficient in FileMaker Pro to effectively analyse the project data.

Another alternative is to purchase a SAS module for metadata management. The SAS Warehouse Administrator software allows database managers to create and manage databases through a single point of control, resulting in improvements in speed, quality and consistency of data delivered to end users but at a cost. The product retails for over $40,000 and is therefore exceeds the budgetary constraints of LSAC.

While STATA has a number of estimators suitable for analysis of longitudinal data, it does not at present have any commands built-in specifically designed for metadata management and analysis. Similarly, SPSS does not offer a metadata management tool within their software products. Of the more complex studies using SPSS for data management above, Microsoft Access has generally been employed in the management of data.

Another attractive option for the LSAC Data Management team is the development of a meta-database with a web-based front end which could be accessed by all users, either by accessing the LSAC website or by loading the complete database to a local computer and using a web-browser to manipulate files. This product would be known as the LSAC 'Data Dictionary' and would comprise:

- Data items and their associated variable names;
- Data items linked across successive waves;
- Identification of the construct being measured by data items;
- Thematic groupings of constructs;
- Thematic grouping across successive waves;
- Rationale for the implementation of each grouping – linked to the LSAC key research questions;
- The ability to search for items by question name/number, theme and variable name.

Such a product would circumvent the problems associated with multiple computer platforms and software choices made by the end user. It would also add functionality in excess of most current longitudinal projects. The advantages of the product would multiply exponentially with each successive wave of data collection in terms of data item mapping, reference and manipulation. This is the currently preferred model for the LSAC project.

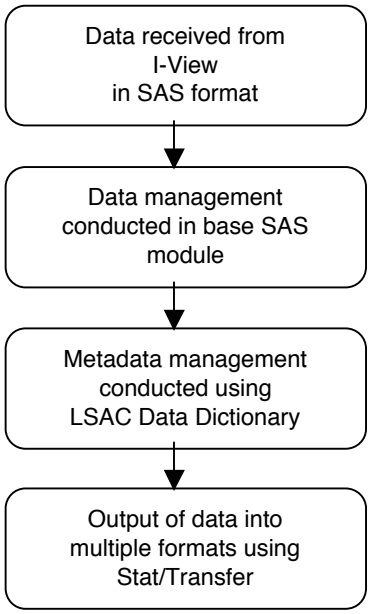
Transferring Files

Amongst local and international researchers, opinion on the most effective means of data transfer between various software platforms is less diffuse. Key respondents identified *STATA's* file transfer program *Stat/Transfer* as being the most effective and efficient software for this task. Output of LSAC data will made available across all major software platforms. STATA's file transfer program accommodates these demands. It handles missing data, value and variable labels and those details that are necessary to move as much information as is possible from one file format to another. Supported formats include Microsoft Access (Versions 2.0 through Office '97), dBase (all Versions), Delimited ASCII, Epi Info, Excel worksheets (all versions, including Excel 2000), Fixed format ASCII, FoxPro, Gauss (Windows and Unix), JMP, LIMDEP, Lotus 1-2-3 (all versions), Matlab 5, Mineset, Minitab, OSIRIS (read-only), Paradox (all versions), Quattro Pro for DOS and Windows, S-Plus (Windows and

Unix), SAS for Unix -- HP, IBM, Sun, SAS for Unix -- DEC Alpha, SAS for Windows and OS/2, SAS Transport, SAS Value Labels, SAS Version 7-9, SPSS Datafiles (Windows and Unix), SPSS Portable Files, Stata (all versions, including 8), Statistica, and Systat (Windows and Macintosh).

Summary

The below figure identifies the software implemented at each stage of data management.



Additional Project Cost:

- Nil

References

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

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

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

Freidin, S., Watson, N. & Wooden, M. (2002), *HILDA Survey Coding Framework: Confidentialised Data,* HILDA project technical paper series, The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

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

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

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

Lepkowski, J.M. (1989) "Treatment of wave nonresponse in panel surveys" in Kasprzyk, D. *et al.*, *Panel Surveys,* Wiley, New York.

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

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

Skinner, C.J. & Holmes, D.J. (2003), "Random effects models for longitudinal survey data", in Chambers, R.L. & Skinner, C.J. eds., *Analysis of Survey Data,* Wiley, Chichester.

Soloff, C., Millward, C., Sanson, A. & the LSAC Consortium Advisory Group (2002), *Proposed Study Design and Wave 1 Data Collection, LSAC Discussion Paper No 2.,* Australian Institute of Family Studies, Mellbourne.

Tabachnick, B.G. & Fidell, L.S. (1989), *Using Multivariate Statistics*, Harper and Row, Sydney.

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

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

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

Wooden, M. & Watson, N. (2000), *The Household Income and Labour Dynamics in Australia (HILDA) Survey: An Introduction to the Proposed Survey Design and Plan,* HILDA project technical paper series, The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

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

Wooden, M. (2001), *Design and Management of a Household Panel Survey: Lessons from the International Experience,* HILDA project discussion paper series, The Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

---

[i] Members of the LSAC Research Consortium who contributed significantly to this paper are: David Lawrence