LONGITUDINAL SURVEYS OF AUSTRALIAN YOUTH

TECHNICAL REPORT 77

Data linkage and statistical matching: options for the Longitudinal Surveys of Australian Youth

Sinan Gemici

Nhi Nguyen

National Centre for Vocational Education Research

Longitudinal Surveys of Australian Youth

NCVER

Longitudinal Surveys of Australian Youth

Data linkage and statistical matching: options for the Longitudinal Surveys of Australian Youth

Sinan Gemici
Nhi Nguyen
National Centre for Vocational Education Research

NATIONAL CENTRE FOR VOCATIONAL EDUCATION RESEARCH
TECHNICAL REPORT 77

Acknowledgment

The authors are grateful to Helen Rogers (Section Manager, Longitudinal Study of Australian Children, Department of Families, Housing, Community Services and Indigenous Affairs [FaHCSIA]) for contributing valuable insights to this discussion paper.

This document should be attributed as Gemici, S & Nguyen, N 2013, *Data linkage and statistical matching: options for the Longitudinal Surveys of Australian Youth*, NCVER, Adelaide.

This work has been produced by NCVER through the Longitudinal Surveys of Australian Youth (LSAY) Program, on behalf of the Australian Government and state and territory governments, with funding provided through the Australian Department of Education, Employment and Workplace Relations.

ISBN 978 1 922056 59 7
TD/TNC 112.18

Published by NCVER, ABN 87 007 967 311

Level 11, 33 King William Street, Adelaide SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

P +61 8 8230 8400 **F** +61 8 8212 3436 **E** lsay@ncver.edu.au **W** <www.ncver.edu.au>

About the research

Data linkage and statistical matching: options for the Longitudinal Surveys of Australian Youth

Sinan Gemici and Nhi Nguyen, NCVER

Recent evaluations of the Longitudinal Surveys of Australian Youth (LSAY) have recommended investigating the potential for combining LSAY data with external data sources as a way to improve the breadth of information in the survey, but without adding respondent burden. Against this backdrop, the purpose of this discussion paper is to investigate the potential for linking data from existing administrative collections to LSAY and to explore the viability of combining data from LSAY and the Longitudinal Study of Australian Children (LSAC).

Key messages

- Linking administrative data from the education, training and health sectors would greatly enhance the ability to explore key drivers of young people's transition outcomes in LSAY without increasing respondent burden.
- The potential benefits are particularly appealing in topic areas that are currently quite limited in LSAY, such as health information, childhood development and early education outcomes. This makes linking the National Assessment Program — Literacy and Numeracy (NAPLAN) and Medicare data to LSAY the most valuable initial option.
- In a further stage, linking data from the Department of Human Services (Centrelink), the Australian Census, and national education and training statistics to LSAY could provide an evidence base for generating insights into the intergenerational impact of disadvantage.
- Although a statistical match between the Longitudinal Study of Australian Children and LSAY is at first sight appealing, given the complementary nature of these two flagship surveys, a closer look reveals a number of methodological obstacles. Research findings from such an amalgamated dataset of 'synthetic' individuals would lack the necessary robustness to inform evidence-based policy.

Overall, strong consideration should be given to concrete plans for linking administrative collections to LSAY, beginning with NAPLAN and Medicare data.

Tom Karmel
Managing Director, NCVER

Contents

Tables and figures

Tables

Figures

Executive summary

Understanding youth transitions requires information on young people's individual background characteristics and the circumstances under which they grow up. In fact, the ability to assemble information about family and community background, physical health and psycho-social development, as well as academic achievement and the broader school environment, into a coherent data stream, from infancy right through to adulthood, is invaluable for developing effective policy settings. However, no single data source in Australia currently provides coverage of young people's developmental trajectories from birth and early childhood, to tertiary education and entry into the labour market.

One option for addressing the lack of life-course data is to link an existing flagship youth survey such as the Longitudinal Surveys of Australian Youth (LSAY) to existing administrative collections, such as the National Assessment Program — Literacy and Numeracy (NAPLAN), Medicare Australia and others. Data linkage refers to the process of matching records on the same person held in different data sources, such that the different sources are combined to present more comprehensive information on individuals. With an increasing number of Australian and international research projects capitalising on the advantages of data linkage, the idea of supplementing LSAY with data from administrative collections is well worth exploring.

Another option for creating life-course data from existing sources is to enhance LSAY with information from the Longitudinal Study of Australian Children (LSAC). While both surveys collect data on background characteristics and key life events, they do so for different sets of individuals and across different age groups. LSAY could be complemented with information from the Longitudinal Study of Australian Children via 'statistical matching', whereby individuals from both surveys who are statistically equivalent on a number of key background characteristics are merged into a fictitious individual to observe the impact of socio-demographic attributes and key interventions on transition outcomes over time.

The purpose of this discussion paper is to evaluate the feasibility of both approaches. In the first part of the report, the potential for enhancing LSAY with information from administrative collections through data linkage is investigated. The second part explores the viability of combining relevant data from LSAY and the Longitudinal Study of Australian Children via statistical matching.

When exploring the possibilities of data linkage with LSAY it is necessary to consider the challenges inherent in the process. These challenges revolve around technical issues, cost, and legal/ethics considerations. The latter generally represent the largest obstacle, given that legal consent needs to be sought from LSAY respondents (and possibly their parents or legal guardians) in order to proceed with any form of data linkage. Privacy regulations further require that specific protocols be followed to ensure the protection of privacy and confidentiality. These regulations include the de-identification of linked data, use of an independent agency as data custodian and integrating authority, and secure storage of linked data. The costs associated with observing privacy regulations would likely be offset by the advantages data linkage can bring to LSAY; namely, broadening the scope of the questionnaire without increasing respondent burden. Existing questions could also be supplemented with administrative data, allowing scope for new questions.

The specific administrative data sources considered for linkage with LSAY in this discussion paper are the National Assessment Program — Literacy and Numeracy, Medicare Australia, the Department of Human Services (Centrelink), the Higher Education Statistics Collection, the National VET Provider Collection and the Australian Census. While all six collections could add considerable value to LSAY, the largest initial benefit would be derived from linking the National Assessment Program — Literacy and Numeracy and Medicare data to LSAY. Centrelink data could then be linked in a second step to further enhance the breadth and depth of LSAY.

The Longitudinal Study of Australian Children provides valuable data on early childhood development (among a host of other relevant information), whereas the Longitudinal Surveys of Australian Youth focus on the transition experience from 15 years of age onwards. If it were feasible to combine relevant information from the Longitudinal Study of Australian Children and LSAY, researchers would be able to analyse a powerful dataset that captures aspects of a person's developmental trajectory from birth up to about 25 years of age.

Given that LSAY and the Longitudinal Study of Australian Children do not contain the same individuals, a method known as 'statistical matching' would have to be employed to combine both surveys. Statistical matching combines records of individuals who are statistically similar on key characteristics and which are available in both datasets. From a conceptual perspective, it is important to understand that combining LSAC data with LSAY via statistical matching would result in a synthetic dataset, in which each matched record represents a fictitious individual who proxies the combined trajectory of two real individuals who are statistically equivalent on a number of key characteristics.

Although statistical matching can be useful in certain situations, this report concludes that it is not advisable to combine information from LSAY and the Longitudinal Study of Australian Children into a synthetic dataset for further analysis. A statistical match between LSAY and the Longitudinal Study of Australian Children is an interesting empirical exercise, yet the methodological obstacles are such that any results from an analysis of a matched Longitudinal Study of Australian Children—LSAY dataset would lack the necessary robustness to inform policy and practice in meaningful ways.

The overall conclusion from this discussion paper is that strong consideration should be given to concrete plans for linking administrative collections to LSAY, beginning with the National Assessment Program — Literacy and Numeracy and Medicare data. Once a process for data linkage has been developed for LSAY, linking with either Centrelink data or the Australian Census could be investigated, based on a detailed cost—benefit analysis.

Introduction

Recent evaluations of the Longitudinal Surveys of Australian Youth (LSAY) have recommended investigating the potential for combining LSAY data with external data sources as a way to improve the breadth of information in the survey without adding respondent burden. The stocktake report on LSAY (Nguyen et al. 2010) identified the potential of linking administrative collections to LSAY as a strategy to improve data quality, as well as to provide more comprehensive information on respondents' family background and health. A recent review of the LSAY program also recommended potentially linking LSAY with appropriate administrative datasets (for example, the National Assessment Program — Literacy and Numeracy [NAPLAN]) and discussed conducting an investigation into the feasibility of combining LSAY with data from the Longitudinal Study of Australian Children (LSAC) as a strategy for investigating early childhood development and experiences during the school-to-work transition.

In response to this context, this paper explores two issues: the potential for linking data from existing administrative collections to LSAY; and the feasibility of combining data from the Longitudinal Study of Australian Children and LSAY.

Background

Understanding youth transitions requires information on young people's individual background characteristics and the circumstances under which they grow up. Such information includes family and community background, physical health and psycho-social development, as well as academic achievement and the broader school environment. The ability to assemble this information into a coherent data stream from infancy through to adulthood is invaluable for developing effective policy settings. In addition to informing policy-makers and practitioners about the need for policy intervention, such comprehensive life-course data can shed light on the question of when different interventions have the strongest positive impact on transition outcomes.

No single data source in Australia currently provides longitudinal data on young people's developmental trajectories from early childhood to tertiary education and entry into the labour market. Australia's two child/youth flagship surveys, the Longitudinal Study of Australian Children and the Longitudinal Surveys of Australian Youth, collect detailed information on background characteristics, educational achievement and key life events for different sets of individuals and across different age groups. Administrative collections such as Medicare Australia and Centrelink,[1] or point-in-time collections such as the Australian Census, also contain important data on factors that directly or indirectly influence young people's transition outcomes. Combining elements of different data sources can potentially generate a coherent data stream that cannot otherwise be gained from a single survey or administrative collection.

[1] Centrelink has been incorporated into the Department of Human Services. For simplicity, the term 'Centrelink' will be used throughout this report.

The Longitudinal Surveys of Australian Youth

Managed and funded by the Australian Government Department of Education, Employment and Workplace Relations (DEEWR), LSAY is a research program that tracks young people as they move from school into further study, work and other destinations. It uses large, nationally representative samples of young people to collect information about education and training, work and social development.

Survey participants in the current LSAY collection enter the study at 15 years of age. Individuals are contacted once a year for up to 12 years. Studies began in 1995 (Y95 cohort), 1998 (Y98 cohort), 2003 (Y03 cohort), 2006 (Y06 cohort) and more recently in 2009 (Y09 cohort). Over 10 000 students start out in each cohort. Since 2003, the initial survey wave has been integrated with the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA). Table 1 provides a brief overview of LSAY characteristics.

Table 1 Overview of LSAY characteristics

Cohort	Initial sample size	Survey period	Age range	Age at most recent available wave
Y95	13 615	1995–2006	15–25	25
Y98	14 117	1998–2009	15–25	25
Y03	10 370	2003–2013	15–25	22
Y06	14 710	2006–2016	15–25	19
Y09	14 251	2009–2019	15–25	16

Note: In the Y95 and Y98 cohorts, the sampling criterion was students in Year 9 rather than students 15 years of age. Therefore, in Y95 and Y98 the average age when first surveyed was 14.7 years.

The LSAY research program provides a rich source of information to enable a better understanding of young people and their transitions from school to post-school destinations; it also explores some social outcomes, such as wellbeing. Information collected as part of the LSAY program covers a wide range of school and post-school topics, including: student achievement, student aspirations, school retention, social background, attitudes to school, work experiences and what students do when they leave school. As part of the Programme for International Student Assessment, from 2003 the base year of each LSAY cohort also provides detailed information about respondents' school environment, as outlined in table 2.

Table 2 Major LSAY topic areas

Student-related	School-related
Demographics (student; parent)	Structure and organisation
Education (school; school transition; post-school)	Staffing and management
Employment (current; job history and training; seeking employment; not in labour force)	Resources
Social (health, living arrangements and finance; general attitudes)	Accountability and admission practices

Linking LSAY to other datasets

Because the main focus of the surveys is tracking youth transitions, LSAY is limited in the areas of health and wellbeing and early childhood development. With an increasing number of Australian and international research projects capitalising on the advantages of data linkage, the idea of supplementing LSAY with data from administrative collections to broaden the scope of the surveys is one worth exploring as a model for future LSAY development.

Linking administrative data to surveys such as LSAY enables researchers to investigate issues that cannot be explored by any single source of information. The lack of contextual information in administrative collections limits researchers who require data on a range of background factors that may be causally related to a given outcome of interest. Surveys such as LSAY capture such information and provide rich data in other areas, such as socio-demographic background, participation in specific school and out-of-school interventions, personal attitudes and aspirations, and the family or neighbourhood environment. However, surveys such as LSAY are constrained by population size, which limits investigation into subgroups and into the impact of policy 'treatments'.

Unlike surveys, which take a population sample, administrative databases usually collect data for all members of a specific population of interest (for example, all people in receipt of unemployment benefits). This feature guarantees large datasets, which in turn improves the precision of statistical analyses (Jenkins & Siedler 2007). Other benefits of administrative data include accuracy and cost-effectiveness. Administrative data can be more accurate when compared with information provided in self-reported surveys because they are less prone to recall error (George & Lee 2002).[2] The use of administrative data can also result in substantial cost savings if variables of interest are available and no further data need to be collected through surveys (Hoffmann 1995).

Despite these advantages, data linkage is not without its challenges. The first part of this paper explores the concept of data linkage and considers advantages/issues of linking LSAY to:

- National Assessment Program — Literacy and Numeracy
- Medicare Australia
- Centrelink
- Higher Education Statistics Collection
- National VET Provider Collection
- Australian Census.

An alternative method for combining datasets is examined in the second part of this paper, where the concept of 'statistical matching' is explored, whereby data from different sources, as opposed to data linkage, which do not contain the same individuals, are combined. Using this methodology, existing LSAY data could be combined with relevant early childhood information from the Longitudinal Study of Australian Children. The final part of the report identifies the most viable option(s) for LSAY.

[2] We should acknowledge, however, that there are also sources of error or potential inaccuracy in administrative data. These include data entry errors (not always subject to the quality control found in survey data entry) and definitional issues (for example, definitions that do not match those used in statistical collections).

Exploring data linkage with LSAY

The concept of data linkage

The concept of data linkage[3] is becoming increasingly popular as a way of combining relevant information from different sources. Data linkage refers to the process of matching records held in different data sources about the same person (Jutte, Roos & Brownell 2011). Data linkage thus only applies to situations in which the different data sources to be combined contain, at least in part, information on the same individuals. For instance, it is feasible to combine records from Medicare Australia with LSAY via data linkage because all LSAY respondents should, in theory, be included in the Medicare database.[4] Data linkage becomes less useful as the population overlap between data sources decreases, and the method is not applicable when attempting to combine data sources that do not contain information about the same individuals. A potentially suitable method for this latter scenario is considered in a later section.

Data are linked via deterministic or probabilistic methods. A combination of linkage methods may be used in any one project, but the choice of method depends on the types and quality of linkage variables available on the datasets to be linked (Australian Institute of Health and Welfare 2012). Deterministic data linkage can be used when records across different data sources have a common unique identifier, also known as a 'statistical linkage key'. Statistical linkage keys across different datasets are found in community service program data collections in Australia, but are rarely available in other administrative data collections, and so probabilistic data linkage is mostly used in practice. The probabilistic method links records on a combination of several high-quality representative identifiers, such as a person's full name, gender, date of birth and address. Representative identifiers are used to compute the probability of two records from different data sources belonging to the same individual. The two records are then linked once a certain probability threshold is reached.

The actual task of identifying and linking records from different sources is carried out automatically using probabilistic linkage algorithms.[5] While complications can arise from duplicate records or typographical errors in the data, modern computer algorithms achieve very accurate linkage results. Jenkins et al. (2008) linked a large British household survey to an administrative tax collection to benchmark probabilistic versus deterministic methods. Probabilistic methods yielded accurate results when linking on gender, date of birth, postcode, and first and family name as representative identifiers. The authors attested that 'record linkage between household survey responses and administrative data is feasible, and … can yield good results when judged in terms of numbers of matches and their accuracy' (Jenkins et al. 2008, p.40).

Use of data linkage in current research

One important driver for the recent proliferation of data-linkage studies in Australia and abroad is the need to understand the complex interactions between intergenerational and early childhood health

[3] 'Data linkage' is also known as 'record linkage', and these terms are sometimes used interchangeably. In this report, the term 'data linkage' is used throughout.

[4] LSAY participants would have Medicare records either on their own account or through their parents or legal guardians. The availability of Medicare records might be limited for recent immigrants to Australia.

[5] Technical details on probabilistic linkage algorithms are provided in Blakely & Salmond (2002) or Tromp et al. (2011).

factors on one hand, and broader social outcomes over an individual's life course on the other. Understanding these relationships requires linking data sources across sectors such as health, education, welfare and other relevant sectors.

Significant work based on linked life-course data has recently been undertaken in Canada. Using linked data from the Manitoba Population Health Research Data Repository,[6] researchers have established important causal relationships between early life risk factors and long-term health, education and labour market outcomes (Brownell et al. 2010; Jutte, Brownell et al. 2010; Jutte, Roos et al. 2010). Similar cross-sectoral studies with linked data have been conducted in Sweden (Lawlor et al. 2006; Li, Sundquist & Sundquist 2010). Current Australian examples of cross-sectoral linkage projects include the prediction of reading and numeracy skills from early childhood development data (Gregory 2012) and the impact of social and clinical background factors on school readiness (Lynch 2012). A brief summary of selected research studies using linked data is provided in table 3.

Table 3 Summary of selected research studies using linked data

Study focus	Author(s)	Location	Data sources
Associations of child socioeconomic status and mortality	Lawlor et al. (2006)	Sweden	• Swedish Multi-generation Register • Swedish Cause of Death Register • Swedish Census
Impact of adolescent socioeconomic status on higher education participation	Chowdry et al. (2008)	UK	• English National Pupil Database • Higher Education Statistics Agency Student Records
Academic and social outcomes for high-risk youth	Brownell et al. (2010)	Canada	• Manitoba Population Health Research Data Repository
Biologic versus social predictors of childhood health and educational outcomes	Jutte, Brownell et al. (2010)	Canada	• Manitoba Population Health Research Data Repository
Social, educational, and medical outcomes for children of teenage mothers	Jutte, Roos et al. (2010)	Canada	• Manitoba Population Health Research Data Repository
Effects of parental occupation on low birth weight	Li et al. (2010)	Sweden	• WomMed II[7] • Swedish Census
Development characteristics at age five as predictors of reading and numeracy skills three to seven years later	Gregory (2012)	Australia	• Australian Early Development Index • Western Australian Literacy and Numeracy Assessment • NAPLAN
Impact of social and clinical background factors on school readiness	Lynch (2012)	Australia	• Births, Deaths and Marriages • Children, Youth and Women's Health Service: Child Health Record (0–4 years) • SA Health: Integrated South Australian Activity Collection, Emergency Department Data Collection, Perinatal Data Collection • SA School Enrolment Census, NAPLAN, Australian Early Development Index

Note: Studies are listed in chronological order. Full citations are provided in the references section.

[6] The Manitoba Population Health Research Data Repository links the population registry to several health, education and welfare databases. For details see University of Manitoba (2012).

[7] WomMed II is a nationwide database that contains information from the Swedish medical birth register, which includes both birth records and prenatal care data (Swedish Centre for Epidemiology 2003). The WomMed II database also contains nationwide individual-level hospital diagnoses and death register data, as well as census data.

The Longitudinal Study of Australian Children provides a successful model of linking administrative data to a flagship longitudinal survey. Conducted in partnership with the Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA), the Australian Institute of Family Studies (AIFS) and the Australian Bureau of Statistics (ABS), the Longitudinal Study of Australian Children (also known as *Growing Up in Australia*) collects data to answer the following broad research questions (see Rogers et al. 2009):

- What are the childhood experiences and conditions that impact on child, adolescent and adult outcomes and on trajectories of development?
- What are the mechanisms underlying linkages and interactions and how do these change over time?
- What factors protect children from events or contexts that increase the risk of poor outcomes?

The Longitudinal Study of Australian Children has recruited approximately 10 000 families across Australia to examine child development and follows two cohorts of children: those who were aged 0—1 years of age in 2004 ('infant cohort') and those who were 4—5 years of age in the year 2004 ('child cohort'). The two cohorts are currently aged 9—10 and 13—14 years and data collection for both is planned to continue to cover the transition into adulthood. Data are collected on a bi-annual basis. Table 4 provides a brief overview of the characteristics of the Longitudinal Study of Australian Children.

Table 4 Overview of Longitudinal Study of Australian Children characteristics

Cohort	Initial sample size	Survey period	Age at most recent survey wave
Infant	5 107	2004 – open	9–10
Child	4 983	2004 – open	13–14

Information on children is complemented with data from parents as well as teachers or carers. The survey covers 14 major topic areas relating to the study child and its family, as outlined in table 5. Further information on the Longitudinal Study of Australian Children is provided in Sanson et al. (2002).

Table 5 Major Longitudinal Study of Australian Children topic areas

Child-related	Family-related
Developmental foundations	Family characteristics
Physical health	Family relationships
Social, emotional, behavioural, psychological characteristics	Family interactions, behaviours and wellbeing
Time use and activities	Parenting
Learning competencies and achievement	Family health
Becoming an adult	Family work characteristics
School environment	Family activities
Peer relationships	Family resources
Work	

Major efforts have been undertaken to link the Longitudinal Study of Australian Children to the following administrative datasets (Soloff et al. 2007):

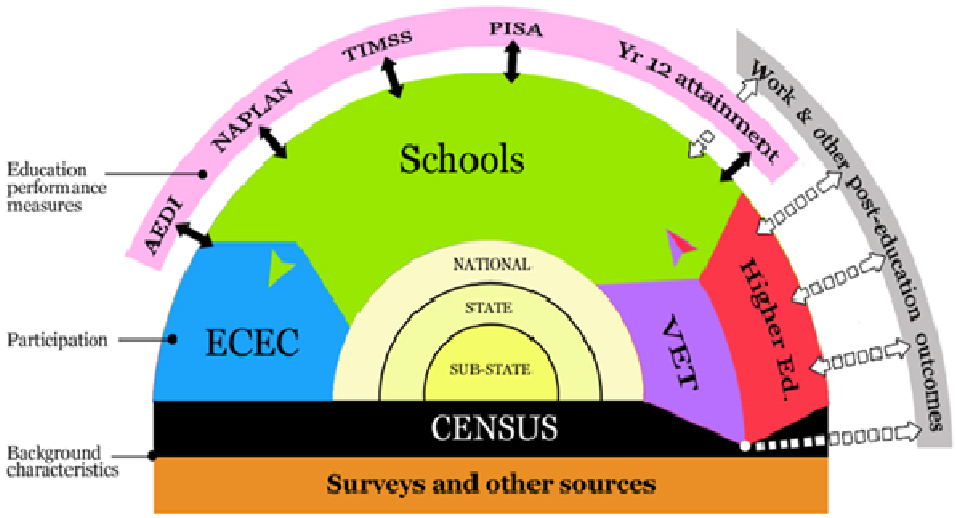
- health and development information recorded in parent-held records about every child after birth
- hospital records of the child's birth

- immunisation records held by the Australian Childhood Immunisation Register
- episodes and types of healthcare utilisation funded by Medicare Australia
- information held by Medicare Australia on the Pharmaceutical Benefits Scheme
- data on the quality of childcare centres and family day care schemes, as held by the National Childcare Accreditation Council
- ABS indices of disadvantage, remoteness indicators and other measures of interest (such as unemployment rates) derived from census data.

Further to the administrative collections listed above, Daraganova, Edwards and Sipthorp (2013) recently illustrated the process of linking NAPLAN academic achievement scores to corresponding LSAC participants, based on a child's first and surname, date of birth and school's name and postcode. The link between the National Assessment Program — Literacy and Numeracy and the Longitudinal Study of Australian Children allows researchers to determine the impact of individual and parental background characteristics, early childhood and school interventions, as well as personal attitudes and aspirations, on academic outcomes in Years 3, 5, 7 and 9.

A concept paper recently presented by the ABS provides an example of how data linkage can help to explore young people's developmental trajectories from early childhood to their mid-20s (National Statistical Service 2012a). The Australian Longitudinal Learning Database (ALLD) project plans to integrate the Australian Census with the National Assessment Program — Literacy and Numeracy and other administrative collections. Data from the Programme for International Student Assessment (and therefore LSAY) and potentially other relevant surveys may also be included. Figure 1 illustrates the various elements that may be included in the Australian Longitudinal Learning Database. However, it needs to be emphasised that this project is currently in very early stages of development and may take years to be fully implemented.

Figure 1 Concept of the Australian Longitudinal Learning Database



Source: National Statistical Service (2012a).

Challenges of data linkage

From a technical viewpoint, it is feasible to combine administrative data with LSAY via probabilistic linkage because respondents' representative identifiers (that is, names, addresses, dates of birth etc.) are known to the data collection contractor (currently Wallis Consulting Group). The data-linkage process itself, however, requires a number of key steps and may vary, depending on the linkage model and the linkage method; that is, whether it is deterministic or probabilistic. Even with best practice models to draw on, significant resources and time will be required to develop a process for linking LSAY to other data collections, including determining the best method for dealing with data-quality issues.

From a legal and ethics viewpoint, the major issue for consideration is protecting privacy and data confidentiality, which includes obtaining consent and dealing with consent bias, and developing specific protocols for the secure management of linked data.

Firstly, any potential data linkage between LSAY and administrative collections will require obtaining consent from respondents to link their data. Because there is no minimum age at which an individual can make decisions on his or her personal information under the *Privacy Act 1988*,[8] LSAY respondents can be asked directly for their consent.[9] This can be practically undertaken through the sample maintenance schedule, whereby respondents are contacted twice a year.[10] A more significant challenge is that of consent bias. Consent bias refers to the fact that population groups with certain socio-demographic characteristics (for example, low socioeconomic status and Indigenous youth, as well as other equity groups) have lower consent probabilities for data linkage (Australian Institute of Health and Welfare & ABS 2012; Kho et al. 2009; Knies, Burton & Sala 2012; Sala, Burton & Knies 2012). This can lead to an under-representation of already vulnerable population groups in statistical analyses that use linked data. In LSAY, consent bias could be addressed to some extent by developing appropriate statistical weights based on consent probabilities. Ideally, consent would need to be obtained at the earliest possible stage to avoid compounding consent bias with attrition. However, the process of obtaining consent may itself have implications for attrition (see Soloff et al. 2007).

Secondly, the development of a data-linkage process using LSAY data needs to take into account specific protocols that ensure the protection of privacy and the confidentiality of the data and manage the safety and security of linked data (National Statistical Service 2010). These include de-identification of linked data, use of an independent agency as data custodian and secure storage of linked data. For data-linkage projects involving Commonwealth data for statistical and research purposes, an official 'integrating authority' must be used. An integrating authority is the single agency ultimately accountable for the implementation of a statistical data linkage project (National Statistical Service 2012b). Currently, the two integrating authorities for Commonwealth data are the

[8] A general principle in determining when a young person has the capacity to make a decision on his or her behalf is when they have sufficient understanding and maturity to understand what is being proposed. In some circumstances it may be appropriate for a parent or guardian to consent on behalf of a young person where the child is very young or lacks the maturity of understanding to do so themselves (Australian Law Reform Commission 2008).

[9] Note that other guidelines adhered to by collection agencies or contractors may set a minimum age requirement. For example, the Australian Market and Social Research Society's (AMSRS) Code of Professional Behaviour, which is followed by the current field work contractor for LSAY, requires researchers to obtain consent from a parent or guardian before a child of 14 years and under can be interviewed. Moreover, some ethics committees may suggest that whether consent from a 15-year-old is acceptable depends on whether it can be reasonably argued that they understand fully what they are consenting to. Best practice would likely seek parental consent in addition to respondent consent until the LSAY respondent reaches the age of 18.

[10] It is generally most appropriate to obtain written consent in addition to the verbal consent given during the telephone interview.

ABS and the Australian Institute of Health and Welfare (AIHW). This means that the actual process of linking any Commonwealth administrative collections to LSAY has to be coordinated by either one of these authorised external intermediaries.

The costs associated with setting up the linkage model and the processes to manage privacy concerns are likely to be offset by the advantages that data linkage brings to LSAY; namely, broadening the scope of the questionnaire without increasing respondent burden. Existing questions could also be supplemented with administrative data, allowing scope for new questions.

Potential administrative data sources for linkage with LSAY

There are numerous administrative collections held at different levels of government which could enhance the breadth of LSAY through data linkage. In identifying the most appropriate datasets for linking with LSAY, the main consideration is whether the additional information outweighs the costs associated with the linkage. Datasets that collect information in areas where LSAY is limited are a natural starting point: information on outcomes prior to age 15 (when LSAY begins) and health information. For these areas, the National Assessment Program — Literacy and Numeracy and Medicare data are two national collections used in previous data-linkage projects that could be explored in a potential LSAY linkage project.

National Assessment Program — Literacy and Numeracy

The National Assessment Program — Literacy and Numeracy was introduced in 2008 to assess all students in Years 3, 5, 7 and 9 in reading, writing, language conventions (spelling, grammar and punctuation) and numeracy. Since 2003, LSAY has also featured cognitive assessment data for 15-year-olds in reading, mathematics and science literacy through its integration with the OECD's Programme for International Student Assessment.[11] The benefit derived by linking NAPLAN scores to LSAY would be access to literacy and numeracy development from Years 3 to 9, allowing researchers to control for academic achievement at earlier ages. Conversely, the lack of contextual information in NAPLAN data could be broadened with individual background and transition data collected from LSAY.

Unlike census data, which are held and managed centrally by the ABS, scores from the National Assessment Program — Literacy and Numeracy are stored by their respective state/territory governments. As mentioned earlier, linking these scores to an existing longitudinal survey has been successfully undertaken, providing a model for linking the National Assessment Program — Literacy and Numeracy to LSAY. As outlined by Daraganova, Edwards and Sipthorp (2013), the steps for linking the National Assessment Program — Literacy and Numeracy to the Longitudinal Study of Australian Children (which would be analogous for LSAY), after obtaining consent and identifying the population to match, include:

- Step 1: as the responsible data-integration authority, the ABS sends each relevant state/territory data authority a list of representative identifiers (for example, first name, surname, date of birth, school name) from LSAY,[12] along with a dummy LSAY ID, which is different from the actual LSAY respondent ID.

[11] Prior to the link between LSAY and PISA from 2003 onwards, the Y95 and Y98 cohorts featured reading and numeracy tests that were administered to students in the respective base year. Test results led to the creation of three school achievement measures: achievement in literacy, achievement in numeracy and combined achievement in literacy and numeracy. Further details are provided in NCVER (2011a).

[12] LSAY respondents' representative identifiers are held by the LSAY data collection contractor (currently Wallis Consulting).

- Step 2: each relevant state/territory data authority matches the National Assessment Program — Literacy and Numeracy and LSAY data on the representative identifiers.
- Step 3: states/territories send the LSAY data collection contractor a list that contains the National Assessment Program — Literacy and Numeracy scores against the dummy LSAY ID.
- Step 4: in order to link the National Assessment Program — Literacy and Numeracy scores to LSAY, the LSAY data collection contractor uses an ABS-generated concordance between the dummy LSAY ID and the actual LSAY respondent ID.

Medicare Australia

Medicare Australia, a Commonwealth Government program, covers a wide range of health care services. Collected primarily to aid with the financing of healthcare, it is among the most accurate sources of healthcare data in Australia (Van Gool, Parkinson & Kenny 2011). Medicare contains data for out-of-hospital services for all patients and in-hospital services for private patients; it does not contain public hospital data, as this information is managed by the states and territories. Within Medicare, collections with particular relevance for data linkage with LSAY include the Medicare claims data and the Pharmaceutical Benefits Scheme (PBS). The Medicare claims data include vital information on the date, location and type of medical service provided. The Pharmaceutical Benefits Scheme ensures that all Australians have access to prescribed medication at a reasonable cost; data from the scheme contain information on the type of drug purchased by an individual, as well as repeat administrations.

Among the major advantages of Medicare data from a research perspective is the accuracy and longitudinal nature of the data. Retrospective data for up to five years at a point in time are available (Medicare Australia can hold data for maximum of five years). Linking Medicare and Pharmaceutical Benefits Scheme claims data to LSAY would allow researchers to gain a more accurate and extensive picture of any medical conditions or health issues that may impact on respondents' outcomes, as this information is currently limited in LSAY. Another benefit is that Medicare data have been widely used in previous data-linkage projects, providing a number of models to draw on.

However, young people in the LSAY age group are relatively low consumers of health care resources, which means the overlap with LSAY may be low. Another issue to consider for the LSAY population is that young people are eligible for their own Medicare card at 15 years of age but they need to enrol in Medicare. However, this issue could be remedied because Medicare can extract young people's data even if they are still on the parental Medicare card.

Other potential administrative data collections

Other administrative data collections that come to mind when considering potential data linkages with LSAY are collections in areas that are currently covered in the survey, but where data linkage would improve the accuracy of the data as well as offer scope to replace existing questions with new fields that capture more complex measures (for example, wellbeing or social capital). However, multiple linkage efforts can be costly and a balance is needed with the number of datasets linking to LSAY. A decision needs to be made about which dataset confers the most benefit to LSAY over and above the opportunity to replace existing questions. The datasets considered in this section include: Centrelink, the Higher Education Statistics Collection, the National VET Provider Collection, and the Australian Census.

Centrelink

Centrelink is a Commonwealth Government program administering income support and government benefits, including parenting and family payments, unemployment benefits, old-age and disability pensions, and student assistance. Centrelink payment records capture history on individual as well as family income support, including payment types and dates of receipt, and provide an opportunity to investigate the consequence of growing up in a welfare-dependent family and the likelihood of a youth's future receipt of income support. The administrative records include whether parents and siblings of an individual have ever received income support or other government benefits (but note that the relationship between 'parents' and 'siblings' is an administrative one and may not be biological [Breunig et al. 2009]). Other details collected include:

- personal details (such as date of birth, sex, country of birth)
- housing details (such as home postcode history, accommodation history, rent type)
- Youth homelessness or independent history (some young adults over 15 years may receive an 'independent status' and be considered independent from their parents for the purposes of Centrelink payments)
- full-time student history
- marital status history.

This administrative information has been previously used for research purposes, most notably as a population sampling frame for the Youth in Focus (YIF) survey. The Youth in Focus surveys followed parents and children from a sample of the Centrelink administrative dataset, with children followed up over two waves (with approximately two years between waves). Respondents were asked for information on employment, education, physical and mental health, attitudes and other psycho-social factors.

Although LSAY captures a great deal of information about a respondent's history of receiving income support, linking with the Centrelink dataset would capture more accurate and detailed information about their payments. For example, in the current surveys there is no distinction made between respondents who receive Youth Allowance or NewStart (unemployment benefits) or whether respondents on Youth Allowance are 'independent' or 'dependent' recipients. More detail about a young person's family history of income support could also be used to better understand the impact of economic and social disadvantage from one generation to the next in identifying, in particular, the most important mechanisms accounting for intergenerational correlation in disadvantage (Breunig et al. 2009).

The main issue is whether data linkage is warranted in the provision of more detailed information about a relatively small population in LSAY. The overlap between individuals in LSAY and Centrelink is relatively low (26%). However, even with relatively small sample overlap, linked LSAY—Centrelink data would provide an invaluable evidence base for generating insights into the intergenerational impact of disadvantage.

Linking Centrelink and LSAY would require consent and linkage procedures similar to those of the administrative datasets previously mentioned, although the linkage with Centrelink is potentially more sensitive and would require parental consent if family records were to be included. Another consideration is the possible risk for attrition from LSAY if respondents feel that the survey is becoming too intrusive and wide-ranging, despite reassurances of confidentiality.

National education and training statistics

One of the major topic areas examined by LSAY is tertiary study, making national education and training statistics likely sources of data for linkage with LSAY. Similar to Centrelink data, these statistics would provide an opportunity to replace existing questions in the survey and have the potential to broaden the information currently collected with additional data about tertiary institutions and subjects and courses. Again, the question is whether these benefits outweigh the associated costs of the data-linkage process. Two key collections to consider are the Higher Education Statistics Collection and the National VET Provider Collection.

Higher Education Statistics Collection

The Higher Education Statistics Collection (HESC) is managed by the Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education (DIICCSTRE) and consists of a comprehensive set of statistics relating to the provision of higher education in all Australian universities. The Higher Education Statistics Collection data relate to:

- courses conducted by higher education institutions
- numbers and characteristics of students undertaking courses
- student load
- completion of units of study and courses
- students' liabilities under the Higher Education Contribution Scheme (HECS)
- numbers and characteristics of staff in higher education institutions
- income and expenditure for higher education institutions
- research activity
- educational profiles of higher education institutions.

Well over 40% of respondents in the LSAY Y03 and Y06 cohorts reported having commenced university study by age 19, providing sufficient overlap between the two datasets to warrant further exploration of data linkage.

Linking relevant data from the Higher Education Statistics Collection may enhance LSAY in several ways. Firstly, LSAY collects data up to the age of about 25 years. This means that information on higher education completion is often not available for young people who begin university study in their early 20s. Linking completion data from the collection could remedy this issue. Secondly, being able to link relevant data from the Higher Education Statistics Collection would provide valuable information on the higher education trajectories and outcomes for those who drop out of the survey. More generally, data linkages such as this provide an opportunity to undertake additional tracking of those who drop out of the LSAY cohort and to determine the extent to which they diverge from those who remain in the survey (that is, whether they are less likely to go on to higher education). LSAY cohorts usually lose over 40% of the initial sample before respondents reach the university entry stage. Moreover, attrition continues after respondents enter university. However, consent to use this information will need to be obtained from these respondents, which may be difficult as they have already ‘disengaged’ from the survey and are thus unlikely to consent. This is a further argument for obtaining consent early in the lifecycle of the survey.

Another benefit to the linkage with HESC data is access to the broader institutional environment in which university study takes place (for example, institutional profiles, staff characteristics,

resourcing). This enables researchers to model multivariate causal relationships between individuals' background characteristics, their secondary schooling experiences and outcomes, their university's environment and institutional characteristics, and higher education outcomes. Such causal modelling can inform the development of interventions for improving higher education uptake and the rates of degree completion. However, is this enough to replace the considerable information relating to higher education that LSAY already collects? LSAY does not collect information on the broader institutional environment at university; however, the Higher Education Statistics Collection data are not as rich as those provided through the Programme for International Student Assessment in relation to information about school institutional factors. Another disadvantage is that Higher Education Statistics Collection data do not provide information on grades or a 'grade point average', which would give researchers a useful measure of 'success' during tertiary studies, one which is not captured in LSAY. Without these additional measures, the Higher Education Statistics Collection data only provide supplementary information, which arguably does not outweigh the costs associated with data linkage. Further, questions about changes to courses, deferral and reasons for not continuing are not captured as they are in LSAY.

National VET Provider Collection

The National VET Provider Collection (also known as the Students and Courses Collection) provides data on the performance of, and outputs from, Australia's vocational education and training (VET) system (NCVER 2011b). Information is collected annually from all public training providers and those private training providers who receive government funding. Records include data on training type and field of study, as well as outcomes for subject and course completions. Additional information includes course delivery type, funding source and recognition of prior learning (RPL). Over a third of Y03 respondents have commenced or completed VET studies by age 19.

The benefits of linking data from the National VET Provider Collection to LSAY parallel those from linkages with the Higher Education Statistics Collection: providing information on subject and course outcomes; possible remedy to attrition from LSAY; and modelling causal relationships between relevant background characteristics and outcomes from VET courses. But the same issues apply: consent to use information on those dropping out of the survey and limited information on the broader institutional environment of the various providers. The latter is subject to more variability than in the Higher Education Statistics Collection data, making analyses more difficult, as identifying 'institutional factors' itself is a topic of ongoing research. Although the VET collection provides information on subject outcomes, 'units of competencies' is not a comparable measure to a student's grade point average: it does not provide an indicator of how well students are progressing by comparison with other students, one that could be used to examine factors related to later employment outcomes. It is important to note that the Apprenticeship and Traineeship Collection could also be considered for data linkage with LSAY, although a smaller proportion of LSAY respondents undertake this pathway.

The main benefit of linking these data collections with LSAY is that they provide an opportunity to replace the current questions in these areas with new questions to capture more complex measures. Because of the resources and time needed to undertake data linkage, it is not feasible to link all of these collections to LSAY. The degree of overlap between the populations in LSAY should be considered, as well as the extent of the additional information that would be useful to researchers and data users.

Australian Census

Managed by the ABS, the census is conducted every five years to measure key characteristics of people and dwellings in Australia. There is a high amount of overlap between the census and LSAY in relation to questions on socio-demographic background, education and employment status. One important exception to this is the provision of neighbourhood data. Researchers agree that neighbourhood attributes have a significant, independent impact on transition outcomes beyond that of individual and family background characteristics (Andrews, Green & Mangan 2004; Cardak & McDonald 2004). Important neighbourhood attributes that can be derived from the census include area-based household income and educational profiles, mobility (that is, the percentage of residents moving in and out of the neighbourhood over a five-year period) and ethnic diversity.[13] Linking these data to LSAY may greatly improve the current understanding of how the neighbourhood environment affects young people's transition outcomes. However, postcode data already collected by LSAY could arguably provide similar information, without the associated costs.

A more technical benefit of linking census data to LSAY lies in addressing the issue of survey attrition. Attrition from the LSAY Y06 cohort, for instance, has been quite high over the first few survey waves (see table 6). Linking data from the 2011 census to this and other cohorts would provide an opportunity to 'backfill' key socio-demographic background data for LSAY respondents who dropped out prior to 2011. However, this option relates more to the sampling design in LSAY rather than being an option for data linkage.

Table 6 Attrition in the LSAY Y06 cohort

Year	2006	2007	2008	2009	2010
No. of respondents	14 170	9 353	8 380	7 299	6 316
% of initial sample	100.0	66.0	59.1	51.5	44.6

[13] For recent applications of census-derived neighbourhood measures in youth research see Jensen & Harris (2008) or Edwards & Bromfield (2009).

Statistical matching

Another method of incorporating external data to broaden the scope of LSAY is statistically matching it to another data source that provides relevant information for exploring the youth transitions currently not captured in LSAY. Because of its focus on early childhood development — unlike LSAY, which captures data on youth transitions from age 15 — one dataset that could potentially be combined with LSAY is the Longitudinal Study of Australian Children. Both surveys share two key objectives:

- to identify factors early on that could increase the risk of negative outcomes in adulthood
- to isolate interventions that could remedy these risk factors.

If it were feasible to add relevant information from the Longitudinal Study of Australian Children to LSAY, researchers would be able to analyse a powerful dataset which captures aspects of a person's developmental trajectory from birth up to about 25 years of age.

The concept of statistical matching

Data from different sources that do not contain the same individuals can be combined via statistical matching (Raessler 2004). Statistical matching combines 'cases that are statistically similar — similar in terms of certain characteristics which are systematically related to the object under investigation and which are observed in a most similar way in both datasets' (Rasner, Frick & Grabka 2010, p.9). From a conceptual perspective, it is important to understand that combining Longitudinal Study of Australian Children data with LSAY via statistical matching would result in a synthetic dataset, in which each matched record represents a fictitious individual who proxies the combined trajectory of two real individuals who are statistically equivalent on a number of key characteristics.

Statistical matching could theoretically be used to combine selected LSAC data with LSAY because both surveys overlap on a number of key matching variables. These matching variables capture mostly time-invariant characteristics such as gender, Indigenous status, socioeconomic status, home language background, location and others. Besides these common variables, each dataset contains unique variables that need to be combined to answer a given research question.

To illustrate this via a highly simplified example: the Longitudinal Study of Australian Children contains detailed information on emotional, physical and social development during childhood. LSAY, on the other hand, contains information on reading, mathematics and science literacy at age 15, as well as Year 12 completion status and post-school engagement in study or work. Table 7 illustrates this simple scenario.

Table 7 Common and unique variables in the Longitudinal Study of Australian Children and LSAY

Common to LSAC & LSAY	Unique to LSAC	Unique to LSAY
Sex	Emotional development	
Indigenous status	Physical development	
Socioeconomic status		Maths/reading/science literacy
Home language background		Year 12 completion status
Location		Post-school study/work status

Note: Adapted from Kiesl & Raessler (2006).

Based on distributional assumptions about the data, statistical matching assigns records in the Longitudinal Study of Australian Children (the donor file) to records in LSAY (the recipient file) for individuals who are statistically similar on key background characteristics. The resulting synthetic dataset would enable researchers to predict literacy scores, Year 12 completion and post-school study/work outcomes, based on emotional and physical development during childhood, while controlling for other relevant confounders. Although this prediction would be for fictitious individuals, their trajectory and outcomes would be representative of people with certain background characteristics of interest. A detailed description of how statistical matching is implemented in practice is provided in the appendix.

Limitations of statistical matching

Several national statistics agencies have explored the viability of statistical matching as a way to combine different datasets in instances where some key variables of interest are not jointly observed (see D'Orazio, Di Zio & Scanu 2003 for Italy; Kiesl & Raessler 2006 for Germany; Liu & Kovacevic 1997 for Canada). Despite this theoretical work, the number of applied research studies that use statistical matching to answer substantial research questions is very limited. One recent example is that of Rasner, Frick and Grabka (2010), who investigated wealth inequalities in the German pension system. Even though relevant administrative information on the same individuals would have been available, data linkage was unfeasible due to privacy legislation. The researchers overcame this issue by using statistical matching as 'the second best solution but the one and only way to overcome the drawbacks of both data sources and make use of their merits' (p.4). Another recent statistical-matching study combined two nationally representative household surveys to analyse the distribution of income and wealth in the United States (Kum & Masterson 2010).

The dearth of applied research studies in which statistical matching is used is due to the fact that the method is based on strong assumptions about the data. These assumptions are often not feasible in practice. The key assumption is that the joint distribution of the two data files is preserved after statistical matching (Raessler 2004). In other words, the resulting synthetic dataset has to adequately capture the relationships between the variables of interest that are not jointly observed in the original datasets. Whether the joint distribution is preserved after matching cannot normally be verified due to the fact that the variables of interest are not jointly observed in the original datasets (the motivation for undertaking statistical matching in the first place).

Distributional requirements are more likely to be met if the assumption of conditional independence holds. Conditional independence requires that the two files to be matched are independent random samples from the same underlying population. Ultimately, however, the validity of a matched synthetic dataset can only be established by checking at least parts of it against a 'true' third source of data. Such a 'true' third data source is rarely available in practice.

Combining LSAY and the Longitudinal Study of Australian Children

The above requirements and assumptions have direct implications for the feasibility of matching data from the Longitudinal Study of Australian Children to LSAY. Given that the two surveys sample individuals from different populations, the conditional independence assumption is violated, which in turn reduces confidence in the joint distribution being preserved. Furthermore, no third source of data is available against which to check the quality of a matched synthetic dataset from the Longitudinal Study of Australian Children and LSAY. An additional issue is that the surveys use

different weights to account for complex sampling and attrition. It is not clear how to statistically handle weights that differ substantially between two datasets (Kum & Masterson 2010).

Overall, while a statistical match between the Longitudinal Study of Australian Children and LSAY is an interesting empirical exercise, the methodological obstacles are such that any results from analysing a matched Longitudinal Study of Australian Children—LSAY dataset would lack the necessary robustness to inform policy and practice in meaningful ways.

Conclusion

This paper has explored two issues: the potential for broadening LSAY with administrative collections through data linkage and the feasibility of combining data from the Longitudinal Study of Australian Children and LSAY via statistical matching.

Statistical matching allows researchers to combine data from different sources about different individuals, based on statistical similarity. Conceptually, a statistical match between the Longitudinal Study of Australian Children and LSAY is at first sight appealing because both surveys share similar objectives and are complementary in terms of the age ranges they cover. In practice, a closer look at matching the two surveys reveals a number of methodological obstacles that greatly reduce the feasibility of a synthetic dataset based on these two surveys. Results generated from such a matched dataset would lack the necessary robustness to inform evidence-based policy.

The alternative option is more viable: broadening LSAY with administrative collections through data linkage. Linked administrative data from the education, training and health sectors would allow researchers to explore key drivers of young people's transition outcomes in LSAY without adding to respondent burden. The benefits are particularly strong in topic areas currently quite limited in LSAY, such as health information, childhood development and early education outcomes. This renders linking the National Assessment Program — Literacy and Numeracy and Medicare data to LSAY the most promising data-linkage option for exploration.

Linking relevant administrative datasets to LSAY also provides an opportunity to generate more accurate information in areas currently collected through respondent recall. This would allow LSAY to focus on the development of the more complex measures driving youth outcomes, such as social capital and wellbeing. The most relevant datasets include Centrelink data, national education and training statistics, and information from the census. The resources and time needed to undertake data linkage means that it is not feasible to link all of these collections to LSAY. The degree of overlap between the populations in LSAY should be considered, as should the extent of additional information that would be useful to researchers and policy-makers.

The number of international research projects taking advantage of data linkage is rapidly increasing, with the Longitudinal Study of Australian Children in Australia providing a successful model of linking administrative data to a flagship longitudinal survey. The conclusion from this discussion paper is that strong consideration should be given to concrete plans for linking administrative collections to LSAY, starting with the National Assessment Program — Literacy and Numeracy and Medicare data. Once a process for data linkage has been developed for LSAY, linking with either Centrelink data or the Australian Census can be investigated, based on a detailed cost—benefit analysis.

References

Andrews, D, Green, C & Mangan, A 2004, 'Spatial inequality in the Australian youth labour market: the role of neighbourhood composition', *Regional Studies*, vol.38, no.1, pp.15—25.

Augurzky, B & Schmidt, CM 2001, *The propensity score: a means to an end*, Working paper no.271, University of Heidelberg, Institute for the Study of Labor.

Australian Institute of Health and Welfare 2012, *Linking SAAP, child protection and juvenile justice data*, AIHW, Canberra.

Australian Institute of Health and Welfare & ABS (Australian Bureau of Statistics) 2012, *National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people*, AIHW, Canberra.

Australian Law Reform Commission 2008, *For your information: Australian privacy law and practice*, ACLR, Canberra.

Baser, O 2006, 'Too much ado about propensity score models? Comparing methods of propensity score matching', *Value Health*, vol.9, no.6, pp.377—85.

Blakely, T & Salmond, C 2002, 'Probabilistic record linkage and a method to calculate the positive predictive value', *International Journal of Epidemiology*, vol.31, no.6, pp.1246—52.

Blakemore, T, Strazdins, L & Gibbings, J 2009, 'Measuring family socioeconomic position', *Australian Social Policy*, vol.1, no.8, pp.121—68.

Bosch-Capblanch, X 2011, 'Harmonisation of variables names prior to conducting statistical analyses with multiple datasets: an automated approach', *BMC Medical Informatics and Decision Making*, vol.11, no.4, p.33.

Brownell, MD, Roos, NP, MacWilliam, L, Leclair, L, Ekuma, O & Fransoo, R 2010, 'Academic and social outcomes for high-risk youths in Manitoba', *Canadian Journal of Education*, vol.33, pp.804—36.

Breunig, R, Cobb-Clark, D, Gørgens, T, Ryan, C & Sartbayeva, A 2009, *User guide to the Youth in Focus data: version 2.0*, Youth in Focus project discussion paper series no.8, Australian National University, Canberra.

Caliendo, M & Kopeinig, S 2005, *Some practical guidance for the implementation of propensity score matching*, Discussion paper no.1588, University of Bonn, Institute for the Study of Labor, Bonn.

Cardak, BA & McDonald, JT 2004, 'Neighbourhood effects, preference heterogeneity and immigrant educational attainment', *Applied Economics*, vol.36, no.6, pp.559—72.

Chowdry, H, Crawford, C, Dearden, L, Goodman, A & Vignoles, A 2008, *Widening participation in higher education: analysis using linked administrative data*, Institute for Fiscal Studies, London.

Daraganova, G, Edwards, B & Sipthorp, M 2013, *Using National Assessment Program — Literacy and Numeracy (NAPLAN) data in the Longitudinal Study of Australian Children (LSAC)*, Department of Families, Housing, Community Services and Indigenous Affairs, Canberra.

D'Orazio, M, Di Zio, M & Scanu, M 2003, *Statistical matching and the likelihood principle: uncertainty and logical constraints*, Italian National Statistical Institute, Rome.

Edwards, B & Bromfield, BM 2009, 'Neighborhood influences on young children's conduct problems and pro-social behavior: evidence from an Australian national sample', *Children and Youth Services Review*, vol.31, no.3, pp.317—24.

Gemici, S, Bednarz, A & Lim, P 2011, *Getting tough on missing data: a boot camp for social science researchers*, NCVER, Adelaide.

George, RM & Lee, BJ, 2002, 'Matching and cleaning administrative data', in *Studies of welfare populations: data collection and research issues*, eds MV Ploeg, RA Moffitt & CF Citro, National Academy Press, Washington, DC.

Gregory, T 2012, 'Early development index (EDI) at age 5 predicts reading and numeracy skills three to seven years later', paper presented at the 2012 International Data Linkage Conference, Perth, viewed 5 December 2012, <http://datalinkage2012.com.au/program/presentations/gregory.pdf>.

Hoffmann, E 1995, 'We must use administrative data for official statistics — but how should we use them?', *Statistical Journal of the United Nations Economic Commission for Europe*, vol.12, pp.41—8.

Heckman, JJ, Ichimura, H & Todd, P 1997, 'Matching as an econometric evaluation estimator: evidence from evaluating a job training programme', *Review of Economic Studies*, vol.64, no.4, pp.605—54.

Jenkins, SP, Lynn, P, Jäckle, A & Sala, E 2008, 'The feasibility of linking household survey and administrative record data: new evidence for Britain', *International Journal of Social Research Methodology*, vol.11, no.4, pp.29—43.

Jenkins, SP & Siedler, T 2007, *Using household panel data to understand the intergenerational transmission of poverty*, University of Essex, Institute for Social and Economic Research.

Jensen, B & Harris, MN 2008, 'Neighbourhood measures: quantifying the effects of neighbourhood externalities', *Economic Record*, vol.84, no.264, pp.68—81.

Jutte, DP, Brownell, M, Roos, NP, Schippers, C, Boyce, WT & Syme, SL 2010, 'Rethinking what is important: biologic versus social predictors of childhood health and educational outcomes', *Epidemiology*, vol.21, no.3, pp.314—23.

Jutte, DP, Roos, LL & Brownell, MD 2011, 'Administrative record linkage as a tool for public health research', *Annual Review of Public Health*, vol.32, pp.91—108.

Jutte, DP, Roos, NP, Brownell, MD, Briggs, G, MacWilliam, L & Roos, LL 2010, 'The ripples of adolescent motherhood: social, educational, and medical outcomes for children of teen and prior teen mothers', *Academic Pediatrics*, vol.10, no.5, pp.293—301.

Kho, ME, Duffett, M, Willison, DJ, Cook, DJ & Brouwers, MC 2009, 'Written informed consent and selection bias in observational studies using medical records: systematic review', *British Medical Journal*, doi:10.1136/bmj.b866.

Kiesl, H & Raessler, S 2006, *How valid can data fusion be?*, Institute for Employment Research of the Federal Employment Services, Nürnberg, Germany, viewed 16 November 2012, <http://doku.iab.de/discussionpapers/2006/dp1506.pdf>.

Knies, G, Burton, J & Sala, E 2012, 'Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population', *BMC Health Services Research*, vol.12, viewed 18 November 2012, <www.biomedcentral.com/content/pdf/1472-6963-12-52.pdf>.

Kum, H & Masterson, TN 2010, 'Statistical matching using propensity scores: theory and application to the analysis of the distribution of income and wealth', *Journal of Economic and Social Measurement*, vol.35, no.3/4, pp.177—96.

Lawlor, DA, Sterne, JAC, Tynelius, P, Smith, GD & Rasmussen, F 2006, 'Association of childhood socioeconomic position with cause-specific mortality in a prospective record linkage study of 1,839,384 individuals', *American Journal of Epidemiology*, vol.164, no.9, pp.907—15.

Leow, C, Marcus, S, Zanutto, E & Boruch, R 2004, 'Effects of advanced course-taking on math and science achievement: addressing selection bias using propensity scores', *American Journal of Evaluation*, vol.25, no.4, pp.461—78.

Li, X, Sundquist, J & Sundquist, K 2010, 'Parental occupation and risk of small for-gestational-age births: a nationwide epidemiological study in Sweden', *Human Reproduction*, vol.25, no.4, pp.1044—50.

Liu, TP & Kovacevic, MS 1997, 'An empirical study on categorically constrained matching', SSC Annual Meeting, Proceedings of the Survey Methods Section, Statistical Society of Canada, Fredericton.

Luellen, J, Shadish, WR & Clark, MH 2005, 'Propensity scores: an introduction and experimental test', *Evaluation Review*, vol.29, no.6, pp.530—58.

Lynch, J 2012, *South Australian early childhood development demonstration project*, viewed 6 November 2012, <www.santdatalink.org.au/Demonstration_Projects>.

National Statistical Service 2010, *High level principles for data integration involving Commonwealth data for statistical and research purposes*, viewed 6 November 2012, <www.nss.gov.au/nss/home.nsf/NSS/7AFDD165E21F34FDCA2577E400195826?opendocument>.

——2012a, *Concept paper: Australian Longitudinal Learning Database (ALLD)*, viewed 6 November 2012, <www.nss.gov.au/nss/home.nsf/NSS/4FB1EC5C8DF5709BCA25784C000386C7?opendocument>.

——2012b, *Rights, responsibilities and roles of integrating authorities*, viewed 6 November 2012, <www.nss.gov.au/nss/home.nsf/pages/Data+Integration+Roles+and+responsibilities+of+integrating+authorities?OpenDocument>.

NCVER (National Centre for Vocational Education Research) 2011a, *Longitudinal Surveys of Australian Youth 1998 cohort: user guide*, NCVER, Adelaide.

——2011b, *Australian vocational education and training statistics explained*, NCVER, Adelaide.

Nguyen, N, Cully, M, Anlezark, A & Dockery, AM 2010, *A stocktake of the Longitudinal Surveys of Australian Youth*, NCVER, Adelaide.

Raessler, S 2004, 'Data fusion: identification problems, validity and multiple imputation', *Austrian Journal of Statistics*, vol.33, no.1/2, pp.153—71.

Rasner, A, Frick, JR & Grabka, MM 2010, *Extending the empirical basis for wealth inequality research using statistical matching of administrative and survey data*, Deutsches Institut fuer Wirtschaftsforschung, Berlin.

Rogers, H, Blakemore, T, Shipley, M & Hutchinson, S 2009, *Longitudinal Study of Australian Children: key research questions*, viewed 8 November 2012, <www.aifs.gov.au/growingup/pubs/reports/krq2009/KeyResearchQuestionsJuly09.pdf>.

Rosenbaum, PR 1998, 'Propensity score', in *Encyclopedia of biostatistics,* eds T Colton & P Armitage, Wiley, New York, pp.3551—5.

Rosenbaum, PR & Rubin, DB 1983, 'The central role of the propensity score in observational studies for causal effects', *Biometrika,* vol.70, no.1, pp.41—55.

——1985, 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score', *The American Statistician*, vol.39, no.1, pp.33—8.

Rubin, DB 1977, 'Assignment to treatment group on the basis of a covariate', *Journal of Educational Statistics*, vol.2, no.1, pp.1—26.

Sala, E, Burton, J & Knies, G 2012, 'Correlates of obtaining informed consent to data linkage: respondent, interview and interviewer characteristics', *Sociological Methods & Research,* vol.41, no.3, pp.414—39.

Sanson, A, Nicholson, J, Ungerer, J, Zubrick, S, Wilson, K, Ainley, J et al. 2002, *Introducing the Longitudinal Study of Australian Children*, viewed 16 November 2012, <www.aifs.gov.au/growingup/pubs/discussion/dp1/discussionpaper1.pdf>.

Schafer, JL 1997, *Analysis of incomplete multivariate data*, Chapman & Hall, Boca Raton.

Smith, JA & Todd, PE 2005, 'Does matching overcome LaLonde's critique of nonexperimental estimators?', *Journal of Econometrics*, vol.125, no.1—2, pp.305—53.

Soloff, C, Sanson, A, Wake, M & Harrison, L 2007, 'Enhancing longitudinal studies by linkage to national databases: growing up in Australia, the Longitudinal Study of Australian Children', *International Journal of Social Research Methodology*, vol.10, no.5, pp.349—63.

Swedish Centre for Epidemiology 2003, *The Swedish medical birth register: a summary of content and quality*, viewed 18 November 2012, <www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/10655/2003-112-3_20031123.pdf>.

Tromp, M, Ravelli, AC, Bonsel, GJ, Hasman, A & Reitsma, JB 2011, 'Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage', *Journal of Clinical Epidemiology*, vol.64, no.5, pp.565—72.

University of Manitoba 2012, *Manitoba population health research data repository list of linked databases*, viewed 18 November 2012, <http://umanitoba.ca/faculties/medicine/units/mchp/protocol/media/database_list_colour.pdf>.

Van Gool, K, Parkinson, B & Kenny, P 2011, *Medicare Australia data for research: an introduction*, Centre for Health Economics Research and Evaluation, Sydney.

Appendix

Using propensity scores to implement statistical matching

Statistical matching is the process of combining records from different data sources on different individuals. The concept of statistical matching is based on combining information from statistically similar individuals from two different datasets. Rasner, Frick and Grabka (2010) implemented the idea of statistical similarity via Mahalanobis distance matching, whereby for each individual x_i in dataset *A* a Mahalanobis distance is calculated to each individual x_j in dataset *B*, based on a number of relevant background variables, or covariates. These covariates are jointly observed in both datasets. The Mahalanobis distance itself measures the distance of individuals from the means of the covariates.

Based on the same idea of statistical similarity, Kum and Masterson (2010) used propensity scores to statistically match two datasets that contain different individuals, a number of jointly observed variables and a set of variables of interest unique to each data source. Their approach is detailed below to illustrate how records from the Longitudinal Study of Australian Children (the donor file) and LSAY (the recipient file) could, in theory, be matched.

Kum and Masterson's (2010) approach for combining two independent surveys is known as propensity score statistical matching (PSSM). The propensity score was originally developed as a means to remedy selection bias in medical studies using observational data (Rubin 1977; Rosenbaum & Rubin 1983). It is based on the concept of randomised control trials and represents the conditional probability of a person being in the treatment group rather than the control group, given a set of relevant covariates.

Within the context of statistical matching, the approach developed by Kum and Masterson (2010) uses the propensity score to determine the likelihood that a record from the donor file can be matched to an equivalent record in the recipient file, based on a number of covariates that are jointly observed in both files. In the context of the present scoping paper, the Longitudinal Study of Australian Children could be considered the donor file and LSAY the recipient file.

A number of steps are necessary to prepare the data files for conducting propensity score statistical matching. Preparation of the data files is a labour-intensive task and entails dealing with missing values and harmonising common variables in terms of their distributions, scales, and ranges.

Missing data

Propensity score statistical matching requires the use of a complete data matrix, and so records with partially missing values have to either be discarded or imputed in some way. When the rate of missing values exceeds 5%, discarding records with missing values is likely to introduce undue bias into statistical procedures (Schafer 1997). In this case, multiple imputation is generally the best alternative to remedy partially incomplete data. For a detailed treatise on handling missing values via multiple imputations in large-scale datasets, readers are referred to Gemici, Bednarz and Lim (2011).

Harmonising common variables

As mentioned above, propensity scores for matching donor file records to those in the recipient file are calculated using a set of jointly observed variables. This requires that each pair of common variables is on the same measurement scale. For instance, if one variable was measured as a continuous variable in the donor file and as an ordered categorical variable in the recipient file, then the continuous variable would have to be converted to a comparable ordered categorical variable. This process is commonly referred to as variable harmonisation (Bosch-Capblanch 2011). To illustrate, family socioeconomic status is measured using socioeconomic position scores derived from parents' educational attainments, their income and their occupational prestige (Blakemore, Strazdins & Gibbings 2009). In LSAY, socioeconomic status is measured using weighted likelihood estimate scores derived from factor analysis on parents' highest educational attainment and occupational status, family wealth, cultural possessions and home educational resources. Statistical matching requires placing both measures of socioeconomic status on an equal footing by standardising both measures. Likewise, harmonising the range of common variables in both data files allows for records to be matched at the margins of the distribution.

Using 'slicing variables'

Slicing variables are used to stratify the datasets to ensure that two matched individuals have sensible characteristics. For instance, gender is a common slicing variable to ensure that records from males in dataset A are matched to records from males in dataset B.

Calculating propensity scores

Propensity scores are calculated using the set of common variables observed in both the donor and the recipient files. The propensity score basically collapses the common set of covariates into a single summary score ranging from *0* to *1*[14]. Formally, the propensity score is expressed as

$$e(x) = pr(z = 1 \mid x)$$

where x denotes the covariates used in the propensity score model, and the binary variable z indicates exposure to treatment (Rosenbaum & Rubin 1985). The propensity $e(x)$ for each individual is estimated through logistic regression of z on x, where z equals *1* for treatment group participants and *0* for control group participants (Rosenbaum 1998). In the context of statistical matching, the propensity scores represent the predicted probability that a donor record will be matched to the recipient file.

Stratification of the data files

Before conducting the match, Kum and Masterson (2010) point out that it is advisable to divide the donor and recipient files into pre-specified strata. The purpose of stratifying the datasets is to

[14] The key benefit of the propensity score is its ability to sidestep the challenges inherent in other matching procedures, which require exact matches on every single covariate in a model (Leow et al. 2004). For instance, if a given multivariate matching model includes 15 dichotomous observed covariates, then 2^{15} (or 32 768) different matches of covariates would be possible. Finding exact matches for treatment participants on all 2^{15} covariates would be an insoluble dilemma, which is known as the 'curse of dimensionality' (Augurzky & Schmidt 2001; Caliendo & Kopeinig 2005). The propensity score overcomes the dimensionality issue by collapsing a large number of observable covariates into a scalar variable between 0 and 1 (Luellen, Shadish & Clark 2005). However, it is important to point out that there is significant variation in matching results based on what particular propensity score matching method is used (for details on this point see Baser 2006).

ensures that specific records are either definitely matched or that matches between specific records are definitely avoided. In the case of matching the Longitudinal Study of Australian Children to LSAY the nature of sub-segments would depend on the research question of interest. For instance, if the analyses of the matched synthetic dataset were to focus on low-socioeconomic status individuals, it would make sense to create a stratum of the lower socioeconomic status quartile in both files and match within this pre-specified stratum of interest.

Conducting the statistical match

The actual process of creating the synthetic dataset is carried out using an algorithm that matches records from the donor file to the recipient file. Basically, the purpose of a matching algorithm is to find the closest match between two records from two different data files in terms of the propensity score distance. With small samples, the choice of one particular algorithm over another can lead to different matching results (Heckman, Ichimura & Todd 1997). With large sample sizes, however, all algorithms perform consistently (Smith & Todd 2005). Given that the Longitudinal Study of Australian Children and LSAY are large-sample surveys, the choice of matching algorithm should thus be inconsequential, unless segmentation leads to small sub-samples.

The concept of matching algorithms is illustrated here using the nearest-neighbour approach, which is a frequently used method in matching studies. The nearest-neighbour algorithm matches a record from the donor file to the most suitable record from the recipient file, based on the proximity of the two records' respective propensity scores. Since it is very rare to find two records from different files with exactly the same propensity score, the nearest-neighbour algorithm seeks to match records whose propensity scores are sufficiently close. Nearest-neighbour matching can be implemented with or without replacement. When matching with replacement, the same record from the donor file can be matched several times to different records from the recipient file. Matching without replacement implies that any treatment group member can be matched with only one member from the control group. Kum and Masterson (2010) strongly recommend matching records without replacement to avoid distorting variable distributions in the matched synthetic file.

One disadvantage of matching without replacement is the possibility of low-quality matches. This can occur, for instance, when for many records with high propensity scores from the donor file there are only few records with high propensity scores in the recipient file. Under this scenario the propensity score distances between two matched records may become unduly large, leading to bad matches. As discussed above, data harmonisation can improve matching quality because it helps to make the joint distributions of the commonly observed variables more similar. Another approach is the use of calipers. Calipers delimit a maximum acceptable propensity score distance so that records from the donor file are matched to those from the recipient file only if their respective propensity scores fall within the specified caliper bounds. While the use of calipers reduces the number of possible matches, matching quality can be greatly improved by avoiding matching records with very dissimilar propensity scores.

Assessing the quality of the matched dataset

The fundamental question that determines the value of a statistically matched synthetic dataset is whether results from statistical analyses closely approach the 'true' results — those that would have been obtained had all variables of interest had been jointly observed. The problem is that, in most practical applications of statistical matching, the 'truth' is unknown. To check the validity of

statistical matching would require verifying that the marginal and joint empirical distributions observed in the matched file approximate those of the donor file. Another source of uncertainty is the significant variation in matching results as a consequence of how exactly the propensity score matching is conducted (Baser 2006). An additional problem specific to matching surveys is that of calculating correct sampling weights; it is currently unclear how to adjust weights correctly in a matched synthetic dataset.

In the end, the difficulties in assessing the quality of a synthetic dataset based on propensity score statistical matching mean that the method is an empirical exercise rather than a viable method for applied research. Kum and Masterson (2010) summarise the limitations inherent in propensity score statistical matching by pointing out that 'if there exists no third source of data against which to check the validity of the synthetic dataset, all that is available in terms of quality control is comparison of the conditional distributions of the donated variables in the donor and synthetic datasets. We acknowledge, of course, that this is a necessary but insufficient indicator of the quality of the match' (p.194).

Australian Government

Department of Education, Employment and Workplace Relations

National Centre for Vocational Education Research Ltd
Level 11, 33 King William Street, Adelaide, South Australia
PO Box 8288, Station Arcade, SA 5000 Australia
Telephone +61 8 8230 8400 Facsimile +61 8 8212 3436
Website www.ncver.edu.au Email ncver@ncver.edu.au