LONGITUDINAL SURVEYS OF AUSTRALIAN YOUTH

TECHNICAL REPORT 86

Linking NAPLAN scores to the Longitudinal Surveys of Australian Youth

MARILYN LUMSDEN
RONNIE SEMO
DAVINIA BLOMBERG
PATRICK LIM

Longitudinal Surveys of Australian Youth

Linking NAPLAN scores to the Longitudinal Surveys of Australian Youth

Marilyn Lumsden
Ronnie Semo
Davinia Blomberg
Patrick Lim

National Centre for Vocational Education Research

LONGITUDINAL SURVEYS
OF AUSTRALIAN YOUTH

TECHNICAL REPORT 86

The views and opinions expressed in this document are those of the author/ project team and do not necessarily reflect the views of the Australian Government, state and territory governments or NCVER.

Any interpretation of data is the responsibility of the author/project team.

This document should be attributed as Lumsden, M, Semo, R, Blomberg, D & Lim, P 2015, *Linking NAPLAN scores to 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 and Training.

COVER IMAGE: GETTY IMAGES/ISTOCK

ISBN 978 1 925173 30 7
TD/TNC 121.15

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

Phone +61 8 8230 8400 **Fax** +61 8 8212 3436
Email ncver@ncver.edu.au **Web** <http://www.ncver.edu.au> <http://www.lsay.edu.au>
Follow us: <http://twitter.com/ncver> <http://www.linkedin.com/company/ncver>

About the research

Linking NAPLAN scores to the Longitudinal Surveys of Australian Youth

Marilyn Lumsden, Ronnie Semo, Davinia Blomberg and Patrick Lim, NCVER

No single data source in Australia currently provides comprehensive longitudinal data on young people's trajectories from early childhood to tertiary education and entry into the labour market. Linking data from the Longitudinal Surveys of Australian Youth (LSAY) with external data sources would improve the breadth of information available from the survey, without adding burden to respondents.

The primary aim of this project is to assess the feasibility (and practicability) of linking National Assessment Program — Literacy and Numeracy (NAPLAN) scores to LSAY data (which contain data from the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA)). A second aim is to determine the similarity between NAPLAN and PISA in measuring underlying academic achievement and whether the two measures rank individuals similarly across the distributions of NAPLAN and PISA.

The NAPLAN tests were first implemented in 2008, which means that the LSAY 2009 commencing cohort (Y09) is the only LSAY cohort to date to have had the opportunity to participate in NAPLAN testing. The analysis undertaken in this paper is restricted to Y09 respondents who participated in the LSAY 2014 survey wave and provided consent to link to NAPLAN.

Key messages

- The project demonstrated that it is technically feasible to link NAPLAN scores to LSAY records; a linking rate of 98% was achieved for consenting LSAY participants.
- It is important to consider more effective strategies to maximise the pool of LSAY respondents available for data linkage. The following strategies are suggested:
  - consider obtaining approvals through existing national governance processes established to support the work of the Commonwealth Government's Education Council rather than separately for each state and territory, with the Commonwealth playing a key role in coordinating changes to the current agreements and existing protocols to support this.
  - obtain consent at the earliest possible time to maximise the number of records available for linking (which also helps to remove bias).
  - avoid the use of written methods in obtaining consent where possible. Telephone and online methods provide better rates of consent.
- The statistical analysis of the NAPLAN and PISA scores showed that there is a reasonable level of agreement between the two measures.
- Expanding the data linkage exercise by joining to multiple years of NAPLAN results would increase the power of the LSAY data by enabling research into the influence of early education outcomes on young people's transitions from school to post-school education and the labour market.

Dr Craig Fowler
Managing Director, NCVER

Contents

Tables and figures

Tables

Figures

Acknowledgments

The authors wish to thank all those who contributed and provided valuable input to this project, including the Australian Bureau of Statistics (ABS), the Australian Institute of Health and Welfare (AIHW) and the South Australian Department for Education and Child Development (DECD). Thanks also go to each state and territory's test administration authority for their generous support, their cooperation in supplying NAPLAN data and providing valuable information that enabled the project's success. Thanks to Dr Sinan Gemici for his methodological advice and work to establish and coordinate this project during the early phases and also to Dr Alice Bednarz for her contribution to the earlier stages of the project. We would like to acknowledge Josephine Foti from the Wallis Consulting Group who was extremely accommodating and supportive throughout the project. We extend our thanks to the numerous reviewers for their considered comments, which have substantially improved the paper. Thanks also go to the Australian Government Department of Education and Training for undertaking the risk assessment and for their ongoing support for the project. The authors also wish to thank Penelope Curtin for her contribution.

Executive summary

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 (Gemici & Nguyen 2013). Linking administrative data from the education, training and health sectors to LSAY data would greatly enhance the ability to explore the key drivers of young people's transition outcomes.

The aim of this project is to assess the feasibility (and practicability) of linking National Assessment Program — Literacy and Numeracy (NAPLAN) scores to LSAY data (which contain data from the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA)). A second aim is to determine the similarity between NAPLAN and PISA in measuring underlying academic achievement and whether the two measures rank individuals similarly across the distributions of NAPLAN and PISA.

LSAY, PISA and NAPLAN

The Longitudinal Surveys of Australian Youth (LSAY) tracks young people as they move from school into further study, work and other destinations using large nationally representative samples of 15-year-olds. Surveys are conducted annually over a ten year period to capture information about young people's transitions from school to tertiary education and the labour market. Since 2003 the initial survey wave has been integrated with PISA.

The Programme for International Student Assessment (PISA) is a triennial international survey that aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students.

National Assessment Program — Literacy and Numeracy (NAPLAN) is the annual assessment of literacy and numeracy performance undertaken by all students in Years 3, 5, 7 and 9. The data from the NAPLAN tests provide schools with information to measure their students' achievements against the national minimum standards.

Many researchers use literacy and numeracy scores from PISA as key predictors of post-school transition outcomes as these scores are available as part of the LSAY dataset. Given that both PISA and NAPLAN scores are routinely used in research studies that inform national education and training policy, it is important to verify that the two measures have a reasonable degree of overlap.

Methodology

The LSAY data are owned by the Australian Government Department of Education and Training and specific arrangements have been established by the Commonwealth Government to manage the risks associated with integrating Commonwealth data. As the custodians of the NAPLAN data, each of the jurisdictions were also required to provide approvals for linking their state or territory's NAPLAN scores to the LSAY data.

In order to link LSAY records to their NAPLAN scores it was necessary to obtain consent from individual LSAY respondents. Three methods for obtaining consent were used — written, oral (via telephone) and online.

The data were analysed through comparisons of summary statistics, graphs and regressions between PISA and NAPLAN to determine the relationship between the two measures.

Findings

The project demonstrated that it is technically feasible to link NAPLAN scores to LSAY records. About four out of five LSAY respondents who had the opportunity to respond to the consent question via their telephone or online interview agreed to have their data linked. We found that obtaining consent using written methods was far less effective, with only one in ten respondents providing consent in this way. Of those providing consent, a matching rate of 98% was achieved overall.

The analysis undertaken in this paper was restricted to a small sub-group of LSAY participants from the 2009 commencing cohort (Y09). The sub-sample comprised those who participated in the 2014 wave of LSAY and provided consent to link to NAPLAN. The analysis showed that this group of participants had higher NAPLAN and PISA scores than the average of all respondents (national average for NAPLAN). The likely reason for this is that higher-performing and more successful individuals are more likely to remain in the LSAY survey over time and may be more likely to provide the required consent to match their NAPLAN and LSAY data.

The secondary purpose of the linkage project was to investigate how similar the PISA and NAPLAN measures are. The statistical analysis showed there is a reasonable level of agreement between the two measures. The weighted correlations were in the range of 0.7 for both maths and reading. The correlations between the NAPLAN reading scores and the PISA reading scores were slightly higher than those for maths.

The future

Despite the high rate achieved when linking the data, it is important to consider how rates of consent can be improved and to develop other strategies to maximise the pool of LSAY respondents available for data linkage. To this end, the following strategies are suggested:

- Avoid where possible the use of written methods in obtaining consent. Telephone and online methods achieve higher rates of consent.
- Obtain consent early to maximise the number of records available for linking, which also helps to remove bias. This could be done by gaining consent during the PISA assessment, or seeking consent during the first round of LSAY interviews.
- Simplify the questions used and information provided during the consent-gathering stage to reduce the burden for interviewers and respondents while ensuring respondents are fully informed.
- Consider obtaining approvals through the existing national governance processes established to support the work of the Commonwealth Government's Education Council rather than separately for each state and territory.

The success in matching NAPLAN scores to the LSAY data means that we can now consider joining multiple years of NAPLAN results. This would allow for the creation of an expanded linked dataset which could be made accessible to researchers and would enable analyses of important policy issues related to the effects of early education outcomes on young people's transition from school to work. Further developments might also include consideration of linkages with other datasets, such as the ABS Census of Population and Housing data (to obtain data on the areas in which respondents live, attend school or undertake further post-school study), and Medicare data.

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, but without adding respondent burden (Gemici & Nguyen 2013). Linking administrative data from the education, training and health sectors to LSAY data would greatly enhance the ability to explore the key drivers of young people's transition outcomes.

Data linkage refers to the process of matching records about the same person held in different data sources (Jutte, Roos & Brownell 2011). Data linkage has been used for health and medical research in Western Australia since the 1970s. In 1995 the Western Australia Data Linkage System (DLS) was established to connect all available health and related information on the Western Australian population to help inform research and projects that aim to improve the health of Western Australians (Data Linkage WA 2015).

At a national level, the Population Health Research Network (PHRN) has been established to build a data-linkage infrastructure for managing health information from around Australia. With the establishment of this network, data-linkage units now operate across every state and territory in Australia. The Population Health Research Network collaboration also involves two national linkage units, namely the Centre for Data Linkage, based in Western Australia, and the Australian Institute of Health and Welfare Data Linkage Unit (Population Health Research Network 2011).

Literacy and numeracy performance in school is a key indicator of how well young people fare after leaving school. The National Assessment Program — Literacy and Numeracy (NAPLAN) is one such measure of literacy and numeracy performance for Australia's school-aged population. Daraganova, Edwards and Sipthorp (2013) recently illustrated the process of linking NAPLAN academic achievement scores to corresponding participants from the Longitudinal Survey of Australian Children (LSAC). The link between NAPLAN and LSAC 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 (Gemici & Nguyen 2013).

Many researchers use literacy and numeracy scores from the Programme for International Student Assessment (PISA) as key predictors of post-school transition outcomes. This is because PISA scores are available as part of LSAY, which tracks nationally representative samples of 15-year-olds for ten years to capture their transition from school to tertiary education and work.

Given that both PISA and NAPLAN scores are routinely used in research studies that inform national education and training policy, it is important to verify that the two measures have a reasonable degree of overlap.

The research objectives of this project are twofold: the first objective is to assess the feasibility of linking NAPLAN scores to LSAY data; the second is to determine the similarity between NAPLAN and PISA in measuring the underlying academic achievement trait and whether the two measures rank individuals in a similar way across the distributions of NAPLAN and PISA.

The LSAY 2009 commencing cohort (Y09) was chosen because it is the only LSAY cohort that has had the opportunity to sit the NAPLAN test, which was introduced nationally in 2008. One jurisdiction was selected to participate in the first stage of this pilot project. For subsequent stages, the remaining jurisdictions were used to top up the initial sample size for the data analysis and compare methods for obtaining consent.

This report begins with an introduction to the datasets used as part of the NAPLAN—LSAY data linkage and the regulations and arrangements that govern their use. The subsequent section describes the process and methods used to undertake this project including: the project approval process; the methods used for obtaining consent; and how the linking of the datasets was undertaken. This is followed by an analysis of the linked dataset to examine the relationship between the two measurements of student achievement. The final section of the report discusses some key considerations emerging from the NAPLAN—LSAY data-linkage experience and investigates what the linkage process might look like when expanded to an entire LSAY cohort.

This research project builds on a preliminary investigation that explored options for linking LSAY data to a range of administrative data. More information on this topic can be found in Gemici and Nguyen's data linkage report (2013).

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 (Gemici & Nguyen 2013).

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, LSAC and LSAY, 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, or point-in-time collections such as the Australian Early Development Census and the ABS Census of Population and Housing, also contain important data on factors that directly or indirectly influence transition outcomes for children and young people. 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 (Gemici & Nguyen 2013).

Longitudinal Surveys of Australian Youth (LSAY)

Managed and funded by the Australian Government Department of Education and Training, 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.

The surveys have a long history and can be traced back to the Youth in Transition (YIT) studies, which began in the late 1970s, with the aim of learning more about the labour market experiences of young people (Karmel 2013). The Australian Longitudinal Survey (ALS) and Australian Youth Survey (AYS) were introduced in the 1980s with a similar remit to the YIT study. These three surveys were combined in 1995 to form the LSAY program.

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), with subsequent cohorts recruited in 1998 (Y98 cohort), 2003 (Y03 cohort), 2006 (Y06 cohort) and, more recently, in 2009 (Y09 cohort). About 14 000 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. Table 1 provides a brief overview of characteristics for each LSAY cohort.

Table 1 Overview of LSAY characteristics

Cohort	Initial sample size	Survey period	Age range	Age at most recent available wave[1]
Y95	13 613	1995–2006	15–25	25
Y98	14 117	1998–2009	15–25	25
Y03	10 370	2003–2013	15–25	25
Y06	14 170	2006–2016	15–25	22
Y09	14 251	2009–2019	15–25	19

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

[1] Refers to the latest survey data available at the time of writing for the two active LSAY cohorts (Y06 and Y09). Data up to the 2013 surveys are publically available.

The LSAY research program provides a rich source of information enabling 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. From 2003, as part of the PISA dataset, the base year of each LSAY cohort includes information about respondents' school environments.

Table 2 outlines the major topic areas covered as part of the LSAY program, further details can be found in appendix A. For more information about the LSAY program visit <http://www.lsay.edu.au>.

Table 2 Major LSAY topic areas

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

Programme for International Student Assessment (PISA)

PISA is a triennial international survey and an initiative of the OECD that aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. To date, students representing more than 70 countries and economies have participated in the assessment.

PISA develops tests that are not linked directly to the school curriculum in the participating countries. The tests are designed to assess the extent to which students can apply their knowledge to real-life situations and are equipped for full participation in society at the end of compulsory education.

Since 2000, 15-year-old students from randomly selected schools worldwide sit the PISA tests in the core domains of reading, mathematical and scientific literacy. Each year of assessment sees a greater focus on one domain.

In an effort to identify the factors that influence student performance and give context to the PISA achievement scores, the PISA student questionnaire collects information on students' backgrounds.

Students are also asked a series of questions about their life at school and the relationship they have with their teachers. Contextual information about the major domain is also collected; this includes attitudes to learning, levels of engagement, activities undertaken and learning strategies used. School management information and instructional practices are also collected as part of the PISA school questionnaire.

The link between LSAY and PISA provides a basis for investigating the enduring effects of the skills, knowledge and other attributes measured in PISA. For more information about PISA, visit <http://www.acer.edu.au/ozpisa>. Further details about the information collected as part of PISA can be found in appendix A.

National Assessment Program – Literacy and Numeracy (NAPLAN)

NAPLAN is an annual assessment undertaken by all Australian students in Years 3, 5, 7 and 9. It tests skills in reading, writing, language conventions (spelling, grammar and punctuation) and numeracy. NAPLAN is the measure through which governments, education authorities, schools, teachers and parents are able to determine whether or not young Australians have the literacy and numeracy skills that provide the critical foundation for other learning and for their productive and rewarding participation in the community.

NAPLAN is one measure of literacy and numeracy performance for Australia's school-aged population. The data from the NAPLAN tests provide schools with information to measure their students' achievements against national minimum standards and student performance in other states and territories.

The administration of the NAPLAN tests is managed by the test administration authority in each state or territory (see table 3). The data resulting from the NAPLAN tests are collected and stored by each jurisdiction's test administration authority, each having its own data-release policies and protocols. Reports on individual student performance are provided to all students and parents/carers by the relevant state or territory authority.

The Australian Curriculum, Assessment and Reporting Authority (ACARA) is the independent authority responsible for developing and managing the National Assessment Program. ACARA also administers the My School website, which reports data from NAPLAN at the school level. The website can be used to view how each year group in a particular school has performed in NAPLAN tests throughout their schooling, including a measure of the gain in student achievement between testing years. The website also provides the capability to compare statistically similar schools and displays school-level information such as staffing, financial information, resources and schools' student characteristics.

For further information on NAPLAN, see <http://www.nap.edu.au/naplan/naplan.html>. Further details on information collected as part of NAPLAN can be found in appendix A.

Table 3 NAPLAN test administration authorities

State/territory	Organisation
New South Wales	Board of Studies, Teaching and Educational Standards
Victoria	Victorian Curriculum and Assessment Authority
Queensland	Queensland Curriculum and Assessment Authority
Western Australia	School Curriculum and Standards Authority
South Australia	Department for Education and Child Development
Tasmania	Department of Education
Northern Territory	Department of Education and Children's Services
Australian Capital Territory	Education and Training Directorate

The linked dataset

This project explores the feasibility of linking data for individuals who were in Year 10 in the LSAY 2009 cohort to their 2008 Year 9 NAPLAN scores. The NAPLAN tests were first implemented in 2008, which means that the LSAY 2009 cohort is the only LSAY cohort to date to have had the opportunity to participate in NAPLAN testing. Year 10 students who participated in PISA in 2009 would have also sat the NAPLAN tests when they were in Year 9 in 2008, as outlined in table 4.

Table 4 Achievement testing and LSAY survey schedule for the LSAY 2009 cohort, 2008–19

Age in years (average)	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
Year level[1]	9	10	11	12								
Year	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*
NAPLAN	✓											
PISA[2]		✓										
LSAY		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: 1 Year level specifies the modal year level for PISA/LSAY participants. NAPLAN assessments are based on year level, while the PISA assessments are age-based. This means that students sitting the NAPLAN test in Year 9 span a range of ages. In contrast, PISA participants are 15 years old when they complete the assessment and span a range of year levels.

2 PISA participants' contact details are collected and used for subsequent interviewing as part of LSAY.

In 2009, those who were in Year 10 in LSAY represented about 70% of all LSAY respondents (see table 5). We further note that the distribution of Year 10 respondents varies between 45% and 85%, dependent upon jurisdiction. This is because the school starting ages differ across the jurisdictions.

Table 5 LSAY respondents by school year level (%), 2009

Year level	NSW	Vic.	Qld	WA	SA	Tas.	NT	ACT	All	All (n)
Year 9	11.3	20.7	1.5	1.3	5.5	33.0	5.4	15.5	**11.3**	**1 617**
Year 10	83.7	77.3	50.5	45.2	84.8	66.9	83.4	83.4	**71.3**	**10 163**
Year 11	5.0	2.0	47.9	53.2	9.6	0.1	11.2	1.1	**17.3**	**2 461**
Year 12	0.0	0.0	0.1	0.3	0.1	0.0	0.1	0.0	**0.1**	**10**
Total (%)	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Total (n)	**3 313**	**2 296**	**2 531**	**1 486**	**1 524**	**1 277**	**788**	**1 036**	**14 251**	**14 251**

Table 6 shows the number of Year 9 students who participated in each of the Year 9 reading and numeracy NAPLAN tests in 2008. Also shown are the:

- total number of participants in PISA 2009 (14 251)
- number of PISA/LSAY respondents who were in Year 10 in 2009 (10 163)

- number of LSAY respondents who were in Year 10 in 2009 and remained in the sample in 2014 (4188).

These 4188 individuals who were in Year 10 in 2009 and remained in the sample in 2014 represent the maximum number of LSAY respondents available to provide their consent to match their NAPLAN scores to their LSAY records. Further information about gaining consent from LSAY respondents to undertake the data linkage can be found in the Methodology section.

The optimal final dataset for this project would include all LSAY Y09 variables (2009—14) and the scaled (overall) reading and numeracy scores for all 4188 individuals from the Year 9 NAPLAN 2008 tests. However, as outlined in the following sections, not all 4188 individuals were contacted because not all of these respondents:

- participated in their 2014 interview
- were able to be contacted within the available timeframes
- gave consent to link their data.

Table 6 Participation in NAPLAN, PISA and LSAY (n)

	Number
Year 9 NAPLAN 2008[1]	
Reading	262 549
Numeracy	262 122
PISA 2009	
Total	14 251
In Year 10 in 2009	10 163
LSAY 2009	
In Year 10 in 2009 and eligible for 2014 LSAY interview	4 188

Note: 1 2008 NAPLAN national report <http://www.nap.edu.au/verve/_resources/2ndStageNationalReport_18Dec_v2.pdf>.

Privacy and data linkage

The Australian *Privacy Act 1988* regulates the handling of personal information about individuals. This includes the collection, use, storage and disclosure of personal information, and access to and correction of that information. The Privacy Act includes 13 Australian Privacy Principles (APPs), which apply to the handling of personal information by most Australian Government agencies and some private sector organisations (Office of the Australian Information Commissioner 2015).

Although identifying information is removed from the LSAY datasets and the data are managed in accordance with privacy protection legislation, it may be possible to use combinations of characteristics (for example, gender, postcode and school type) to re-identify individuals. When linking two datasets, the risk of identification is increased. For this reason, datasets that contain identifiable information need to be handled with care to protect the identity of an individual or organisation.

In 2010, Australian Government Portfolio Secretaries endorsed seven high-level principles for data integration as well as a supporting set of governance and institutional arrangements. If a data-linking project involves Commonwealth datasets and is for statistical and research purposes the project should comply with the *High level principles for data integration involving Commonwealth data for statistical and research purposes* (National Statistical Service 2010).

The contact details of LSAY respondents are handled in the strictest confidence in accordance with the privacy principles. Details about any individual in the LSAY surveys are never made available in LSAY reports or elsewhere. The names and contact details for every LSAY participant are kept in a secure database by the fieldwork contractor (Wallis Consulting Group) and these details are stored separately from the data collected during the annual interviews. Data management and analysis are undertaken by NCVER; the data files contain no contact details. The de-identified datasets are also made available to researchers and other users. The datasets are deposited with the Australian Data Archive (ADA) at the Australian National University in Canberra, and permission to use the data and access requirements are managed by the archive.

NAPLAN test data (which include students' contact details) are collected and stored by each jurisdictions' test administration authority, each with its own data-release policies and protocols. Each state and territory policy seeks to protect personal information and respect the interests of individuals, schools and education agencies. Accordingly, the state and territory test administration authorities will not supply information that identifies or may identify a student, school or sector without the express consent of the student (or parent/guardian in the case of minors).

Undertaking a data-linkage exercise between LSAY and administrative collections such as NAPLAN requires obtaining consent from individuals to link their data. Information about the methods used to gain consent and the consent-gathering process can be found in the Methodology section of this report.

Methodology

The LSAY data are owned by the Australian Government Department of Education and Training. Specific arrangements for linking Commonwealth data were proposed in 2010 to manage the risks associated with integration and to encourage Commonwealth agencies to share their data for linking purposes in a safe and effective way. Consistent and robust processes were proposed to increase Commonwealth agencies' confidence in data-integration projects, in particular the management of systemic risk (National Statistical Service 2013b).

The Commonwealth arrangements were not mandatory for this project, at the time being applied only to a group of selected Commonwealth-approved projects. Nevertheless, the project team followed the National Statistical Service guidance for data-integration projects by completing a risk assessment in consultation with the LSAY data custodian. Further information about the risk assessment is given in the section 'Risk assessment'.

In its capacity as an accredited data-integration authority, the Australian Institute of Health and Welfare (AIHW) was consulted as part of this project. As the custodians of the NAPLAN data, each jurisdiction was also required to provide project approvals for linking their state or territory's NAPLAN scores to the LSAY data. All jurisdictions had an internal approval process which assessed the project methodology prior to providing approval.

Before providing project approvals, at least one jurisdiction required assurances that the project would not involve any analysis or reporting of results by school sector. This requirement was adopted for the project overall (that is, no analysis was conducted by school sector for any state or territory).

A summary of the agencies involved in the LSAY—NAPLAN data linkage project are outlined in table 7.

Table 7 Agencies consulted as part of the LSAY—NAPLAN data-linkage project

Organisation	Role
Program level	
Australian Government Department of Education and Training	LSAY data custodian
National Centre for Vocational Education Research	LSAY data management and analysis
Wallis Consulting Group	LSAY fieldwork contractor
National level	
Australian Institute of Health and Welfare	Advisory role
National Statistical Service	Authority on data-integration projects; risk assessment
State/territory test administration authorities	
Board of Studies, Teaching and Educational Standards	NAPLAN data custodian (New South Wales)
Victorian Curriculum and Assessment Authority	NAPLAN data custodian (Victoria)
Queensland Curriculum and Assessment Authority	NAPLAN data custodian (Queensland))
School Curriculum and Standards Authority	NAPLAN data custodian (Western Australia)
Department for Education and Child Development	NAPLAN data custodian (South Australia)
Department of Education	NAPLAN data custodian (Tasmania)
Department of Education and Children's Services	NAPLAN data custodian (Northern Territory)
Education and Training Directorate	NAPLAN data custodian (Australian Capital Territory)

Risk assessment

A formal risk assessment was undertaken to help assess the level of risk of this project. It was understood at the outset that data-integration projects involving Commonwealth data must undergo a risk assessment, as outlined in the data-integration risk assessment guidelines (see National Statistical Service 2013b). The purpose of the risk assessment is to help Commonwealth agencies assess the level of risk of data-integration projects as part of determining whether a project should proceed and whether an accredited integrating authority is required to manage the integration project (National Statistical Service 2013b). However, as noted earlier, the Commonwealth arrangements were only specifically applied to a group of selected projects and were otherwise not mandatory.

Although the risk assessment was no longer required for this project, it proved useful because it:

- demonstrated good practice in terms of the 'gold standard' for data-integration projects
- assisted the jurisdictions with their internal approval processes
- could be a requirement in the future.

The project was given a 'low risk' rating. The allocated rating indicated that an accredited integrating authority was not required to manage the project.

Further details about the risk assessment process can be found in appendix B or by referring to *Data integration involving Commonwealth data for statistical and research purposes: risk assessment guidelines* available at: <http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/59fd060543b4e9e0ca257a4e001eacfe/$FILE/Risk%20Assessment%20Guidelines%20-%20December%202013.pdf>.

Consent

In order to undertake a data linkage between LSAY and administrative collections such as NAPLAN, obtaining consent from individuals to link their data is required.

Three consent approaches were used — written, oral (via telephone) and online. Oral consent was obtained as part of the annual 2014 LSAY interviews via computer-assisted telephone interview (CATI) for respondents who completed their 2014 interviews by phone. Online consent was obtained as part of the annual 2014 LSAY interviews via computer-assisted web interview (CAWI) for respondents who completed their 2014 interviews online.

The use of three different approaches for gaining consent allowed for some testing and evaluation of the different consent-gathering approaches.

Further information about the methods and guidelines used to ensure that consent had been acquired appropriately can be found in appendix C along with information about the state and territory requirements for gaining consent. The text used to obtain consent for the different approaches can be found in appendix D.

Fieldwork

LSAY respondents are contacted annually up to the age of 25 using telephone interviews. Since 2013, respondents have had the opportunity to complete their interviews online. The fieldwork period commences in July and concludes in January of the following year, and offers an opportunity to request consent to undertake data linkage from respondents.

NCVER's decision to obtain consent using the three approaches of written, telephone and online was prompted by two issues: the jurisdictions' differing consent-obtaining requirements; and the differences in the timing of project approvals from the various jurisdictions. The consent-gathering process therefore involved three sequential stages.

- *Stage one*: the first stage involved the collection of written consent from respondents from the first jurisdiction selected to participate in the project prior to the annual LSAY interviews. This took place during a four-week period from May to June 2014.
- *Stage two*: for the second stage, telephone or online consent was gained for one of the jurisdictions approached early in the data-linkage phase of the project.[1] The consent method employed depended on how respondents completed their 2014 interview (that is, telephone or online). This took place during a 12-week period from July to October 2014.
- *Stage three*: the third stage involved the collection of online consent for the remaining six jurisdictions. Online consent was sought over a ten-week period from August to October 2014. [2,3]

Table 8 provides the methods used and timeframes for obtaining consent. Note that the timeframe for gaining consent did not cover the entire LSAY fieldwork period, and so not all LSAY participants had an opportunity to provide their consent (see appendix C). In addition, some respondents were unable to provide their consent because of the method used to complete their interview. For example, in some instances, consent was obtained using online methods only, but the respondent completed their survey by telephone interview. Table 10 shows that fewer than half of eligible LSAY participants had the opportunity to provide consent.

Table 8 Timeframes for obtaining consent

Method used to obtain consent	Number of jurisdictions	Timeframe	Fieldwork undertaken
Stage 1 – written	One	Four weeks	May – June 2014
Stage 2 – online or telephone	One	Twelve weeks	July – October 2014
Stage 3 – online	Six	Ten weeks[1]	August – October 2014[1]

Note: 1 Consent was sought from one jurisdiction during a 12-week period from July to October 2014.

For stage 1 (written consent), consent could be obtained from all those who sat NAPLAN in 2008 (that is, were in Year 10 in 2009) and were eligible to be interviewed in 2014. For stages 2 and 3 (telephone/online), consent could only be obtained for those who sat NAPLAN in 2008 (that is, were in Year 10 in 2009) and completed their 2014 interviews. The total number of respondents able to provide their consent is summarised in table 9.

[1] This jurisdiction was one of the first jurisdictions approached to take part in the project. As a result, project approvals were obtained for this jurisdiction towards the beginning of the project. For this reason, respondents from this jurisdiction were given priority in terms of fielding their interviews to allow for more respondents to provide their consent.

[2] Consent was sought for one jurisdiction at the very beginning of the fieldwork (that is, from July 2014).

[3] Respondents undertaking their annual interviews at the very beginning of the fieldwork (from July to August 2014) were asked whether they would be interested in providing their consent 'in principle'; the consent question was not asked for this group.

Table 9 Number of respondents eligible to provide consent

Method used to obtain consent	Eligible for interview in 2014	Completed 2014 interview	Eligible to provide consent
Stage 1 – written	806	657	806
Stage 2 – online or telephone	554	481	481
Stage 3 – online	2 828	2 513	2 513
Total	**4 188**	**3 651**	**3 800**

Note: Includes those who were in Year 10 in 2009.

Table 10 shows that the best rates of consent are obtained via telephone interview (89%). Rates of consent via the online interview also fared reasonably well (73%). In contrast, rates of consent are particularly low when obtained via written methods (10%).

Written consent rates may have been improved by extending the timeframe further or conducting follow-up telephone reminders; however, this would significantly add to the overall cost.

Table 11 shows the proportion of consenting respondents by selected characteristics. The data show no apparent differences between the rates of consent for males and females; rates of consent were higher for females for some methods, but not for others. Rates of consent were slightly higher for respondents from metropolitan locations than non-metropolitan locations, but these differences were not large.

Table 10 Number of eligible respondents

	Asked				Not asked	Total
Method used to obtain consent	Provided consent	Did not provide consent	Total	Provided consent (%)		
Stage 1 – written	81	725[1]	806	10.0	**	806
Stage 2 – online or telephone	351	94	445	78.9	36[2]	481
Telephone	*232*	*30*	*262*	*88.5*	**	**
Online	*119*	*64*	*183*	*65.0*	**	**
Stage 3 – online	300	93	393	76.3	2 120[2,3,4]	2 513
Total	**732**	**912**	**1 644**	**44.5**	**2 156**	**3 800**

Notes: ** Not applicable

1 Includes two respondents who returned their signed written consent form after the deadline; it does not differentiate between those who did not return the form and those who did not receive the form.

2 Includes respondents who completed their interview after the time period designated for gathering consent (October 2014 – January 2015).

3 Includes respondents who completed their interview at the beginning of the fieldwork period (July – August 2014) with the exception of one jurisdiction where consent was sought during this period.

4 Includes respondents who completed their interview via telephone.

Table 12 shows average (mean) scores in the PISA maths and reading tests for the group of respondents who gave consent alongside the group of respondents who did not give consent. The data show that respondents who provided their consent had higher achievement scores than those who did not regardless of the method used to obtain that consent.

Table 12 also shows that respondents who provided their consent had a higher socioeconomic status (using the PISA index of economic, social and cultural status) than those who did not provide their consent.

Table 11 Proportion of consenting respondents by selected demographics

	Method used to obtain consent		
	Stage 1 – written	Stage 2 – telephone and online	Stage 3 – online
Sex			
Male	9.7	75.6	82.0
Female	10.3	81.7	73.2
Geographic location			
Metro	10.7	79.7	77.7
Non-metro	7.0	75.6	73.7
All	**10.0**	**78.9**	**76.3**

Note: Geographic location refers to the geographic location of the school the respondent attended when they sat PISA in 2009.

Table 12 Average achievement scores and socioeconomic index, LSAY 2009 cohort, 2014

	Mean score		
	Provided consent	Did not provide consent	Total
PISA math score			
Stage 1 – written	576	543	546
Stage 2 – online or telephone	547	520	542
Stage 3 – online	579	540	570
Total	**563**	**540**	**550**
PISA reading score			
Stage 1 – written	589	550	553
Stage 2 – online or telephone	551	522	545
Stage 3 – online	591	546	580
Total	**572**	**546**	**558**
Socioeconomic index			
Stage 1 – written	0.555	0.455	0.465
Stage 2 – online or telephone	0.459	0.407	0.448
Stage 3 – online	0.598	0.413	0.555
Total	**0.527**	**0.446**	**0.482**

Note: The socioeconomic index refers to the PISA index for economic, social and cultural status.

Linking the data

Data linkage refers to the process of matching records about the same person held in different data sources (Jutte, Roos & Brownell 2011). Data may be linked via deterministic or probabilistic methods.

Deterministic linking involves the exact matching of information on different records across the datasets being combined for a linking project (National Statistical Service 2013a). The probabilistic method links records on a combination of several high-quality representative identifiers that are used to compute the probability of two records from different data sources belonging to the same individual (Gemici & Nguyen 2013).

Deterministic linking has been used for this project because a series of identifiers available on both the LSAY and NAPLAN datasets can be combined to uniquely identify sample members and thereby facilitate the exact matching of records.

Deterministic linking

The simplest form of deterministic linking uses a unique identifier, such as an Australian Business Number (ABN) or a social security number, to determine whether the records refer to the same entity. If a unique identifier is not available, it is possible to instead select a series of variables that are available on each of the datasets being linked.

For this project, LSAY and NAPLAN records were linked using the following PISA variables:[1]

- first name
- last name
- gender
- month of birth[2]
- year of birth
- school name[3]
- school postcode.

Individuals with missing data on the linking variables were excluded from the linking process.

Separation principle

The separation principle is a mechanism for protecting the identities of individuals and organisations in datasets and is applied as part of the linking process used to form a linked dataset. The separation principle means that no one working with the data can view both the linking (identifying) information (such as name, address, date of birth or school name) in combination with the analysis data (such as tertiary entrance scores, health data or employment status) in a linked dataset (National Statistical Service 2013b).

Under the separation principle, individuals only have access to the information needed to perform their role. Those involved in linking the datasets only see the identifying information needed to create the links between different datasets (such as name and address), while those involved in analysing the integrated data only have access to de-identified data specific to the project requirements.

Figure 1 outlines how a unique linking identifier was used to implement the separation principle for this project. This unique linking identifier:

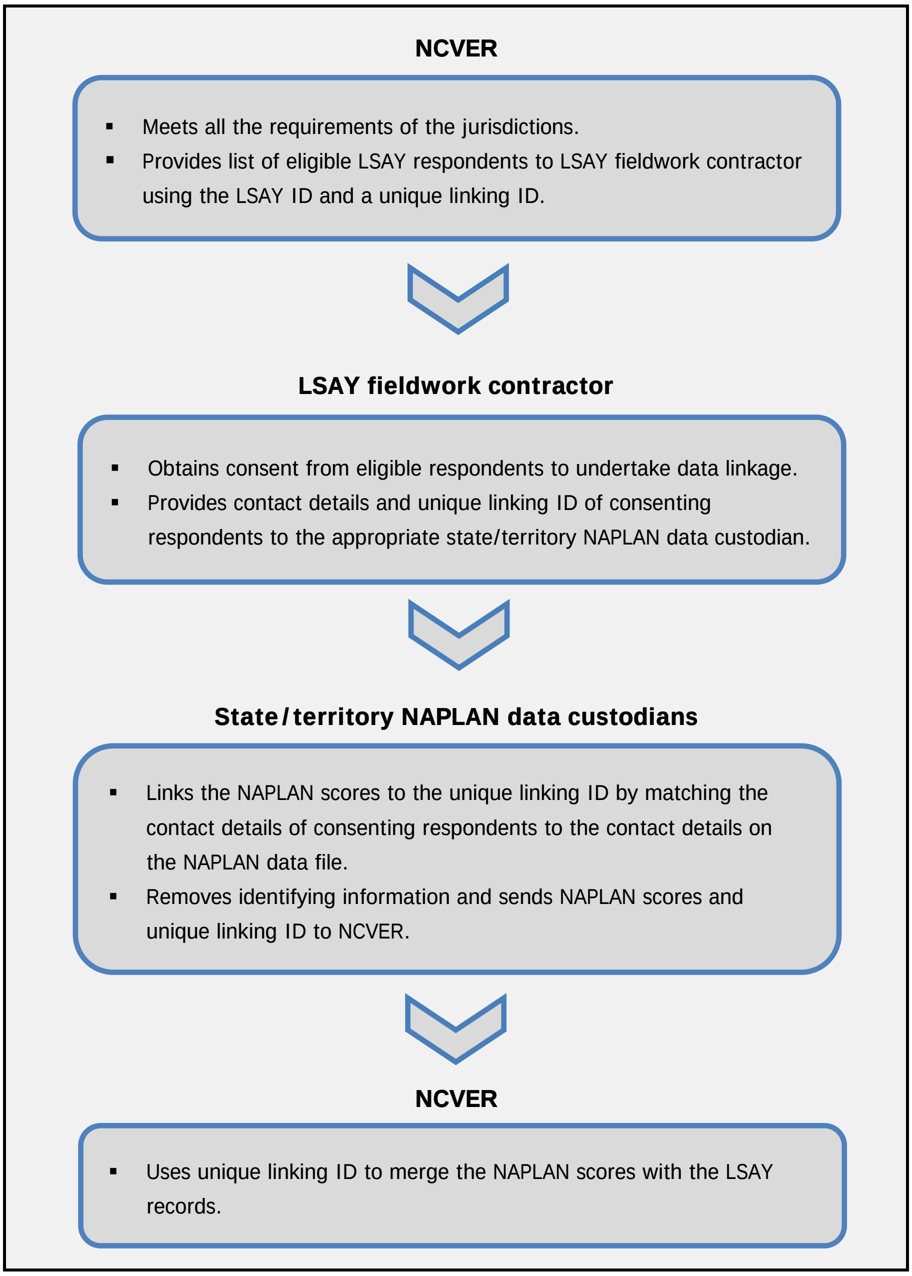
- allows NCVER to merge the LSAY records with the NAPLAN scores (provided by the jurisdictions) without having access to any identifying information
- ensures the state and territory test administration authorities cannot match their own records to the LSAY data.

[1] These variables were provided to each jurisdiction for linkage purposes, but there may be some variation in the variables used to undertake the data linkage.

[2] Date of birth is not available on the PISA dataset.

[3] School name is held by the LSAY fieldwork contractor but is not available on the publicly available PISA dataset.

Figure 1 Data linkage process



NCVER

- Meets all the requirements of the jurisdictions.
- Provides list of eligible LSAY respondents to LSAY fieldwork contractor using the LSAY ID and a unique linking ID.

LSAY fieldwork contractor

- Obtains consent from eligible respondents to undertake data linkage.
- Provides contact details and unique linking ID of consenting respondents to the appropriate state/territory NAPLAN data custodian.

State/territory NAPLAN data custodians

- Links the NAPLAN scores to the unique linking ID by matching the contact details of consenting respondents to the contact details on the NAPLAN data file.
- Removes identifying information and sends NAPLAN scores and unique linking ID to NCVER.

NCVER

- Uses unique linking ID to merge the NAPLAN scores with the LSAY records.

Our approach

In order to link the LSAY 2009 cohort data with the Year 9 NAPLAN scores from 2008, seven steps were undertaken:

1 NCVER contacted each state/territory test administration authority to determine their specific requirements in order to release student-level NAPLAN data.

2 NCVER provided the fieldwork contractor with a list of all eligible LSAY respondents (that is, all LSAY respondents who were in Year 10 when they sat PISA in 2009) using the LSAY identifier and a unique linking identifier.

3 LSAY fieldwork contractor sought consent to undertake the linkage from eligible LSAY respondents.

4 LSAY fieldwork contractor provided the respondent contact details and identifying variables of consenting respondents to the appropriate state/territory test administration authority, along with the unique linking identifier.

5 Each state and territory authority used the contact details and identifying variables provided by the LSAY fieldwork contractor to match the LSAY contact details to the appropriate NAPLAN records.

6 Each state and territory authority provided the de-identified NAPLAN scores and the unique linking identifier for all successfully matched records to NCVER for analysis. All identifying information was removed from the linked files prior to transfer to NCVER. The linked file contains three variables: scaled NAPLAN score for reading; scaled NAPLAN score for numeracy; and the unique linking identifier.

7 NCVER used the unique linking identifier to link the NAPLAN scores received from the jurisdictions to the appropriate LSAY records. NCVER's final file now contains individual-level data (including the LSAY identifier, unique linking identifier, LSAY records and NAPLAN scaled scores) but does not contain any identifiable information.

Table 13 shows the number of respondents from participating jurisdictions[1] who consented to having their data linked and the number of successfully linked records. From the table it can be seen that the matching rates are extremely high, with a 98% matching rate overall. The most likely reason records could not be matched was because respondents had changed schools and/or moved interstate between the time of sitting the NAPLAN test in 2008 and sitting PISA in 2009. It was not possible to compare the characteristics of respondents for which there were matched and unmatched data because the number of unmatched records was so small.

Table 13 Number of records successfully linked

Method	**Consenting respondents (n)**			**Linkage rate (%)**
	Linked	Not linked	Total	Linked
Stage 1 – written	77	4	81	95.1
Stage 2 – online or telephone	344	7	351	98.0
Stage 3 – online	252	2	254	99.2
Total	**673**	**13**	**686**	**98.1**

Note: Includes five jurisdictions who undertook the data-matching exercise. Three jurisdictions were unable to participate as not all requirements or processes were able to be completed within the timeframes available. This equates to 46 respondents who had provided their consent but could not have their records matched for the reasons stated above.

Of the 673 respondents whose LSAY contact details were successfully matched to their NAPLAN records, table 14 shows that 16 respondents were missing a NAPLAN score for numeracy, reading or both. This was primarily because these respondents were absent on the day of the NAPLAN test. A total of 657 respondents were therefore able to have their LSAY records linked to their NAPLAN numeracy and reading scores.

Table 14 Number of records successfully linked by availability of NAPLAN scores

NAPLAN numeracy score	**NAPLAN reading score**		
	Available	Missing	Total
Available	657	5	662
Missing	4	7	11
Total	**661**	**12**	**673**

[1] Five of the eight jurisdictions were ultimately able to participate in the project. Three jurisdictions were unable to participate because not all the specific requirements or processes could be completed within the timeframes or with the resources available.

Project timeframes

The time taken to undertake all the requirements of this project, from understanding the requirements of the project through to assembling the final linked dataset, was lengthy, particularly considering the time taken to obtain approvals and meet jurisdictional requirements.

In future, the time taken to undertake a similar process would be reduced, given the lessons learned and the experience gained from undertaking this project for the first time.

A Gantt chart outlining the time taken to complete each phase of the project can be found in appendix E.

Analysis

NAPLAN is an annual assessment for students in Years 3, 5, 7 and 9. It tests the skills essential for every child to progress through school and life, such as reading, writing, spelling, grammar and numeracy. The tests are constructed to give students an opportunity to demonstrate skills they have learned over time through the school curriculum.

In contrast, PISA assesses the extent to which students near the end of compulsory education have acquired the key knowledge and skills that are essential for full participation in modern societies. PISA is unique because it develops tests which are not designed to be directly linked to school curricula.

Given these differences and that both PISA and NAPLAN scores are routinely used in research studies to inform national education and training policy, it is important to verify the extent to which students' performance in PISA correlates with performance in NAPLAN.

This section presents a descriptive analysis and comparison of the NAPLAN and PISA scores for the matched individuals from the 2009 LSAY cohort. The section will present summary statistics, graphs and simple linear regressions between PISA and NAPLAN to determine the relationship between the two measurements. In particular, the NAPLAN numeracy and reading scores are compared with the PISA maths and reading domains.

Due to the methodology used in obtaining consent, not all LSAY respondents were able to have their NAPLAN data matched. In particular, the final sample contains 657 matched individuals who had both NAPLAN and PISA scores (from a possible 3800 LSAY respondents eligible to provide their consent in 2014). A possible side effect of this is bias. Bias occurs when the sample of interest does not represent the underlying population under consideration. The first step in this analysis therefore is to determine the presence of bias in the LSAY—NAPLAN sample.

The NAPLAN and PISA scores have different underlying distributions - NAPLAN scores are centred at 600 with a standard deviation of around 60, whereas PISA scores are centred at 500 with a standard deviation of 100.

Bias

In order to assess the impact of both attrition from LSAY and non-consent, the distributions of the NAPLAN and PISA scores are presented in table 15 and figure 2.

Table 15 presents the average PISA and NAPLAN scores for the:

- overall population in 2008—09 (column 1)
- matched LSAY—NAPLAN group in 2014 (unweighted in column 2 and weighted in column 3)
- total LSAY sample in 2014 (column 4).

From table 15 we see that the matched group of individuals have both higher NAPLAN and higher PISA scores than the overall population results. The likely reason for this is non-response (attrition) and the fact that higher-performing and more successful individuals are more likely to remain in the LSAY program over time and may be more likely to provide the required consent to match their NAPLAN and LSAY data. The use of weights in LSAY is an attempt to overcome the selection and attrition bias.

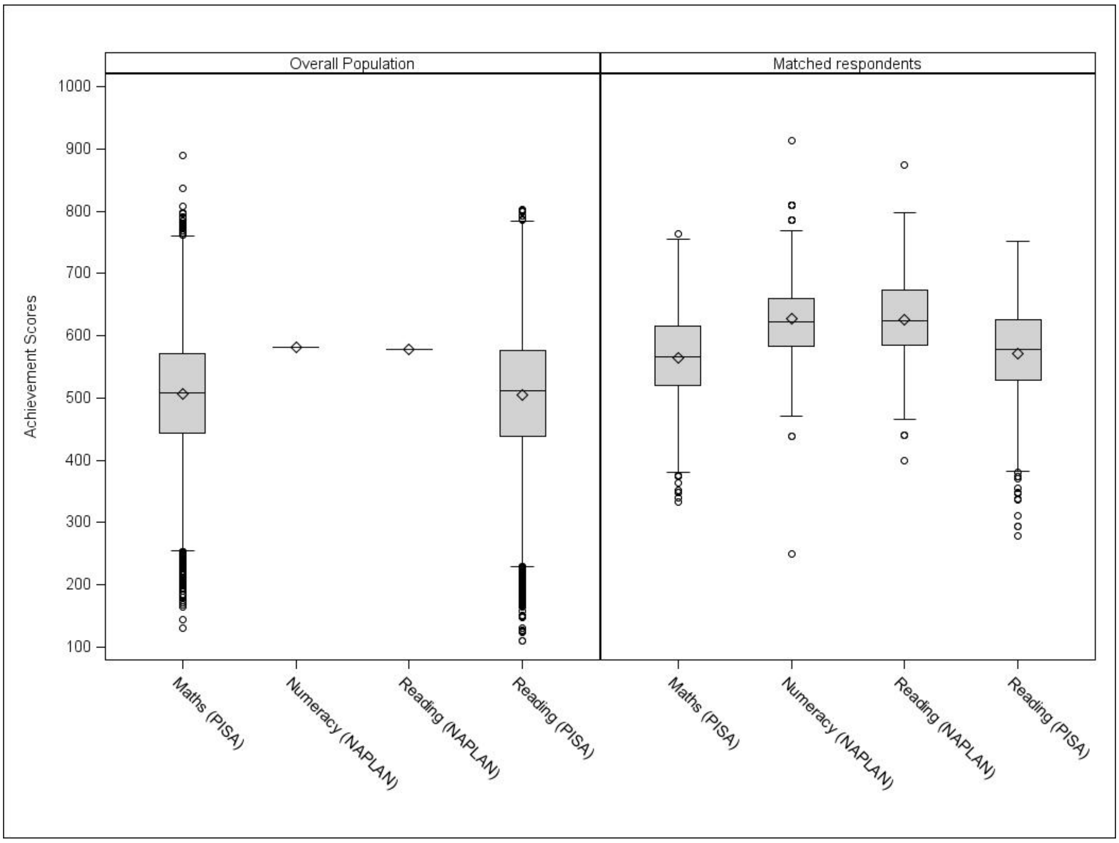
Table 15 NAPLAN and PISA means, weighted

Achievement/NAPLAN scores	Overall population scores	Matched sample, unweighted (2014)	Matched sample, weighted (2014)	LSAY sample, weighted (2014)
PISA maths, 2009	514.55	563.72	548.22	518.63
PISA reading, 2009	514.82	571.66	557.68	524.55
NAPLAN Year 9 numeracy, 2008	582.20	626.86	621.32	**
NAPLAN Year 9 reading, 2008	578.00	625.97	618.85	**

Note: ** Not applicable.

Figure 2 expands on table 15 and presents the mean scores for NAPLAN for all NAPLAN participants and the distribution of PISA scores for all PISA participants (left-hand side of graph), alongside the distribution of scores for those who had NAPLAN and LSAY data matched (right-hand side of the graph). This figure shows that the PISA scores for the overall LSAY sample are more variable than those for the matched respondents. The figure also again shows that the matched respondents have higher average means across the four achievement variables.

Figure 2 Box plots of PISA and NAPLAN scores, unweighted



Typically, analysis using LSAY data uses weighted data and this will correct some of the bias. However, the process for obtaining consent resulted in further bias, and due to the exploratory nature of this project, this bias cannot be adequately addressed by using weighting or another methodology.

As can be seen from table 15, the sub-sample of 657 individuals is still substantially upward-biased, even when appropriate LSAY weights have been applied (column 3), particularly when compared with the weighted scores for the entire 2014 LSAY sample (column 4). The results show that for both PISA and NAPLAN the average scores are higher for the matched samples, and there is also reduced variance for PISA. Thus, it is important to acknowledge that the results that follow — investigating the performance of PISA and NAPLAN — cannot necessarily be extended to the more general population of young people. Given this limitation, it is difficult to assess the relationship between PISA and NAPLAN for those who fall in the bottom end of the achievement distribution.

Relationship

The secondary purpose of the linkage project is to investigate how similar the PISA and NAPLAN measures are. One purpose for determining how the two measures perform in identifying the underlying academic performance distribution is to use the changes in performance in NAPLAN for an individual over time. The relationships presented in figure 3 provide an insight into how well the variables relate to each other; however, future research utilising the linked LSAY/NAPLAN data should investigate whether NAPLAN and PISA predict educational and employment outcomes in a similar way.

Using the unweighted data, figure 3 presents a scatter plot of the NAPLAN numeracy and reading scores against the PISA maths and reading scores. From this figure we can see that there is some agreement between the measures (that is, there is an underlying positive linear trend). From figure 3 it appears as though the relationship between the reading variables is stronger than that of the maths variables. There are three instances where an individual's PISA and NAPLAN scores are markedly different - from the figure we observe that there are two individuals in the maths variables who are outliers from the overall trend and one in the reading who has a higher NAPLAN score than expected, given their PISA score. Therefore, these three individuals are outliers to the overall trend.

Figure 3 Scatterplots of PISA and NAPLAN scores

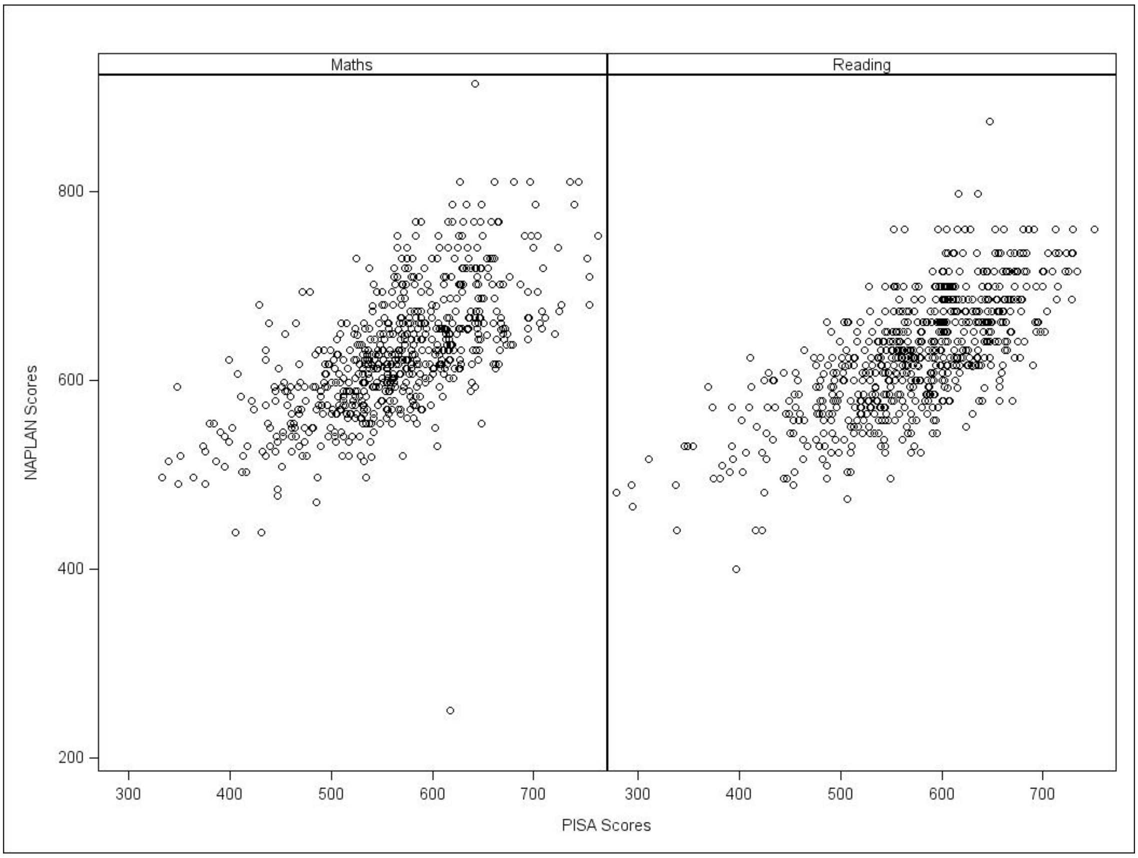


Table 16 presents the correlations between the NAPLAN and PISA scores. It is clear from this table that there is a positive correlation between NAPLAN and PISA. The correlations are of the order of 0.70 (noting that all correlations are statistically significantly (not presented)). We consider the correlation coefficient to be small if its absolute value is less than or equal to 0.3, medium if its absolute value is more than 0.3 but less than or equal to 0.5, and large if it is more than 0.5 in magnitude (Cohen 1988). Thus, we can conclude that there are large correlations between the NAPLAN and PISA scores, indicating that for this particular group of young people individuals fall in similar locations on the distribution for both PISA and NAPLAN.

Table 16 Correlation between NAPLAN and PISA, weighted

Domain	Correlation
Numeracy/maths	0.70
Reading	0.76

As a further demonstration of this, two hierarchical (weighted) regressions were undertaken. The first is that of NAPLAN numeracy against PISA maths scores, and the second the corresponding regression of reading results. The use of a hierarchical model is important because of the nature of the PISA (and NAPLAN) testing. In LSAY, schools are sampled and then individuals within schools are sampled. This sampling structure means that there is likely to be less variance among students within schools than between schools, which needs to be taken into account. A similar result is likely to be observed for NAPLAN. The mixed-model regression results appear in tables 17 and 18 for maths and reading.

Table 17 Regression results, NAPLAN numeracy vs PISA maths

Variable	Estimate	DF	SE	t-value	Pr > |t|
Intercept	317.51	181	14.71	21.58	<0.0001
PISA maths	0.55	476	0.03	21.20	<0.0001
σ^2_{school}	555.25	σ^2_{error}	1564.11	Intra-class correlation	0.26
R-Square	0.51	$\sigma^2_{Total\ (int-only)}$	4358.16		

Note: R-Square value is calculated in mixed models (Nakagawa & Schielzeth 2012).

From table 17, we note that there is a significant positive relationship between the NAPLAN numeracy scores and the PISA maths domain. The PISA maths scores explain around 50% of the variation in the NAPLAN numeracy scores.

The final regression equation to predict the NAPLAN numeracy score from PISA maths is:

NAPLAN numeracy = 317.51 + 0.55 × PISA maths

Table 18 Regression results, NAPLAN reading vs PISA reading

Variable	Estimate	DF	SE	t-value	Pr > |t|
Intercept	316.56	181	11.58	27.33	<0.0001
PISA reading	0.54	472	0.02	26.76	<0.0001
σ^2_{school}	258.02	σ^2_{error}	1224.95	Intra-class correlation	0.17
R-Square	0.59	$\sigma^2_{Total\ (int-only)}$	3656.69		

Note: R-Square value is calculated in mixed models (Nakagawa & Schielzeth 2012).

We see similar results for the reading domains and, as for the correlations, we note that the relationship between NAPLAN and PISA reading scores is slightly stronger than that observed for maths (tables 17 and 18). The regression equation for reading is:

NAPLAN reading = 316.56 + 0.54 × PISA reading

The analysis shows that there is a positive correlation between PISA and NAPLAN with the relationship between the reading variables stronger than the relationship for maths scores.

We note that the benefit of linking of PISA and NAPLAN is not about one measure predicting the other, but whether the PISA and NAPLAN scores can be used interchangeably to predict later education and employment outcomes. A more important benefit in linking LSAY to NAPLAN is being able to access the NAPLAN results for Years 3, 5, 7, and 9 for an individual and determining how the changes in individual performance across the years impact on later educational and employment outcomes. A secondary feature of linking to NAPLAN would be the ability to use the NAPLAN scores in lieu of PISA scores to rebuild the LSAY sample to address attrition. Without the linkage to NAPLAN, there is no means of obtaining a measure of academic performance for new entrants to LSAY.

Discussion

This section discusses some key considerations emerging from the LSAY—NAPLAN data-linkage experience.

Project approvals

Coordinating the differing requirements of each jurisdiction proved to be one of the most challenging aspects of the project. To help remedy this, project approvals could be obtained through existing national governance processes established to support the work of the Education Council, rather than separately for each state and territory, with the Commonwealth playing a key role in coordinating changes to the current agreements and existing protocols to support this.

School sector approvals

To undertake the data linkage for individuals from the non-government school sectors, one jurisdiction required formal approval from each participating independent school and from the Catholic education school body in that jurisdiction. With additional organisations involved, a further step was added to the methodology.

Consent-gathering methods for respondents

The method used to gather consent from respondents to link their LSAY records to their NAPLAN scores had a large effect on consent rates. This result is as expected, given written response methods typically have poorer response rates (Howieson, Croxford & Howart 2008). Gaining consent via telephone interview had the best rates with almost nine in ten respondents providing their consent via this method. Gathering consent via online methods provided good rates with about 73% providing their consent in this way. In contrast, only 10% of respondents provided their consent via written methods.

This poor rate for written consent is likely to have been affected by the timeframe available (four weeks, compared with ten to 12 weeks for telephone and online). Rates of written consent could have been improved by extending the timeframe further and/or by following up respondents by telephone; however these approaches would have added to the cost and timing of the project.

Obtaining written consent is also a more challenging method as respondents need to make much more effort; this includes completing the form, placing it in an envelope and mailing it back to the fieldwork contractor. Providing consent via telephone or online is much simpler as it forms part of a task that respondents are already undertaking as they are required to make a decision at the time of the interview.

The cost of obtaining written consent (which included printing, postage, a reply-paid envelope and follow-up reminder emails to respondents) proved to be considerably higher than the cost of the other methods. This is because both telephone and online consent were added to the end of the annual interviews and the costs were absorbed by the cost of the interviews.

Given that written consent does not yield high response rates, it would be preferable if approval to obtain consent via telephone or online methods is deemed acceptable for all jurisdictions. At the time of the project two jurisdictions required written consent for the release of their NAPLAN records.

However, for one jurisdiction, the existing Act[1] where this is specified is currently under review. It is therefore important that jurisdictional requirements be monitored as these may change in time. For jurisdictions requiring written consent, the LSAY interviews could be used to encourage respondents to complete their consent form.

Respondent feedback

Feedback regarding the consent question (for those who were asked to provide consent via their telephone interview) was obtained from the interviewers during the debriefing conducted by the fieldwork contractor at the beginning of the fieldwork period. The interviewers reported that the consent question was positively received by most responding participants. Interviewers did comment that the four conditions of permission read out by the interviewer were very long for both the interviewer and the respondent. Often the interviewer stated that the respondent required a brief explanation of what was read out, as the information was too detailed for the respondent to absorb.

Any future consent-gathering exercise should attempt to reduce the burden for interviewers and respondents by reducing (where possible) the length and/or complexity of the questions used and information provided. Any simplification to the consent question must still ensure respondents fully understand what they are consenting to. To make sure respondents are well informed about LSAY data-linkage projects and processes, further information should continue to be made available via the LSAY website.

Timing for gathering consent

More LSAY respondent records would be available for data linkage if consent was gained at the earliest possible point in time. This could be done by gaining consent during the PISA assessment, or seeking consent during the first round of LSAY interviews. This would improve the value of the data by maximising the number of respondents asked for consent and increasing the number of linked records.

Obtaining consent at the earliest point in time would also help to remove the bias resulting from obtaining consent from only those who continue to be interviewed in the years following the initial survey wave. Consideration should therefore be given to obtaining consent at the outset of any future commencing LSAY cohort.

Gaining consent as part of the PISA assessment would require coordination with the managers and administrators of PISA, and any changes to the PISA procedures would require collaboration with the governing bodies of PISA. Seeking consent as part of subsequent LSAY interviews may instead prove to be a more pragmatic solution.

Integrating consent to undertake data linkage as part of PISA may raise other issues since PISA participants are only 15 years of age. While the Privacy Act does not specify an age after which individuals can make their own privacy decisions, the Australian Privacy Principles guidelines suggest that an individual aged 15 years or over has capacity to consent unless there is something to suggest otherwise (Office of the Australian Information Commissioner 2014). Despite this, some states and territories may prevent consent from being obtained for participants under 18 years, in which case parental consent would be needed. Countries can choose to take part in a parental questionnaire as part of PISA and this could be an option for obtaining consent for participants under the age of 18

[1] To maintain the confidentiality of participating jurisdictions, specific state/territory acts have not been named.

years. However, it is worth noting that Australia has not opted to participate in the parent questionnaire component of PISA since 2000 and very few countries do. In 2009, only 14 of the 65 participating countries and economies participating in PISA administered the parental questionnaire with varied response rates (Borgonovi & Guillermo 2012).

Linking the data

Linking the NAPLAN and LSAY records (including PISA results) has been extremely successful with a linkage rate of about 98% of consenting individuals. Despite these excellent results, ways in which to improve these linkage rates can be considered.

To aid in the linking exercise, the complete date of birth[1] and whether the respondent has changed schools and/or states (between sitting their NAPLAN test in Year 9 and completing their PISA assessment) could be gathered directly from respondents.

Efficiencies could also be gained by cleaning and standardising the school names on the LSAY and NAPLAN records. One jurisdiction reported inconsistencies in the way in which the school names were recorded when matching the LSAY records to their own NAPLAN records, causing delays in the matching process. Cleaning the school name field on the datasets would help to remove any errors or inconsistencies and would ensure that the fields on each dataset are comparable. This would help to fast-track the matching exercise. Alternatively (or in addition) the school identifiers[2] used by the jurisdictions could be added to the LSAY dataset by the fieldwork contractor to further assist with the matching.

Project timeframes

It is important to appreciate the considerable time required for completing all elements of this project. The time taken to carry out the project, from understanding the requirements of the project through to assembling the final linked dataset was considerable, particularly the time expended obtaining approvals and meeting jurisdictional requirements.

In future, the time required would be reduced, given the lessons learned and experience gained. The length of the process will depend on the method of consent used and whether a consistent approach between jurisdictions is feasible.

[1] Date of birth is not available on the PISA or LSAY data file, only the month and year.

[2] The existing school identifiers on the LSAY dataset are randomised to ensure schools cannot be identified, thereby maintaining confidentiality of the schools participating in the PISA assessment.

Conclusions

The project demonstrated that it is indeed technically feasible to link NAPLAN scores to LSAY records. For respondents who were given an opportunity to provide their consent via their annual LSAY interview about four in every five agreed to have their data linked. Linking rates for consenting respondents were particularly high with a rate of 98% achieved overall.

With such a high linking rate it is important to focus our attention on how rates of consent can be improved as well as develop other strategies to maximise the pool of LSAY respondents available for data linkage.

Obtaining consent at the earliest point in time would improve the value of the data by maximising the number of respondents being asked for consent and increasing the number of linked records. This would also help to remove the bias resulting from obtaining consent from only those who continue to be interviewed in the years following the initial survey wave. In future, telephone and online methods should be used given these provided superior rates of consent, particulary when compared to written methods.

This project was undertaken on the basis of obtaining informed consent from participants. An alternative approach is to undertake data linkage without consent. For a range of government data collections data can be made available for statistical and research purposes, provided there are strong safeguards to ensure the data used for analysis is rendered anonymous or 'de-identified'. A network of state, territory and national data-linkage units has been established in Australia to help facilitate research of this kind. Data linked in this way also reduces selection bias, which may exist if the participants who are successfully contacted and provide consent differ in important ways from those who do not.

The analysis undertaken in this paper is restricted to a small sub-group of participants from the LSAY 2009 commencing cohort, and comprised those who participated in the 2014 survey wave and provided consent to link to NAPLAN. The analysis showed that this group of participants had higher NAPLAN and PISA scores than the average of all respondents (national average for NAPLAN). This limitation made it difficult to assess the relationship between PISA and NAPLAN for those who fall towards the bottom end of the achievement distribution.

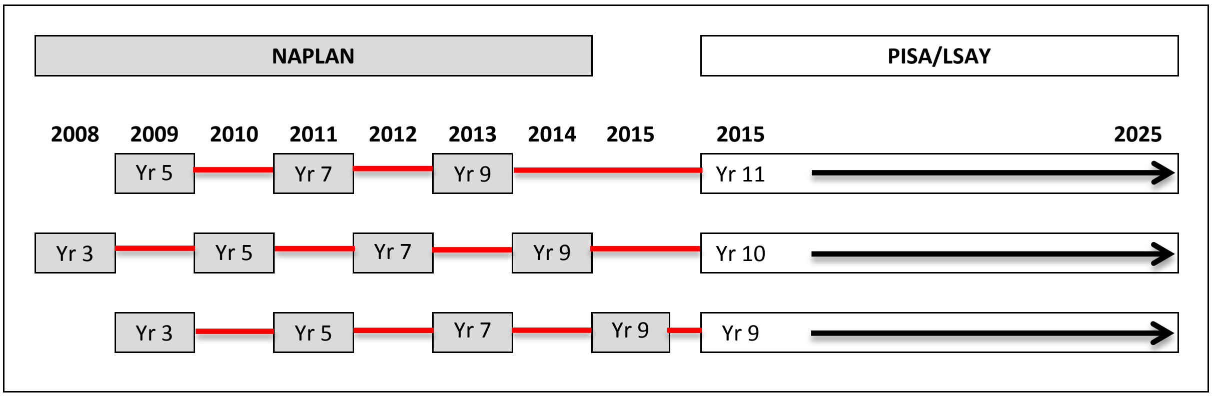
The statistical analysis of the NAPLAN and PISA scores showed that there is a reasonable level of agreement between the two measures. The weighted correlations were in the range of 0.7 for both the maths and reading domains. The correlations between the NAPLAN reading scores and the PISA reading scores were slightly higher than those for the maths domain.

Linking NAPLAN to PISA through LSAY is not concerned with equating one measure with another — it will provide greater insight into the relationship between academic ability and later educational and employment outcomes, given that the two measures are attempting to measure the underlying latent trait of academic performance. We are not proposing that PISA and NAPLAN be linked for the purpose of undertaking comparisons between the two; we are emphasising that they are both measures of the same underlying complex trait and that their linking can be utilised to better inform policy and research questions on youth transitions.

The successful linkage of NAPLAN scores to the LSAY data means that we can now consider expanding the data-linkage exercise by joining multiple years of NAPLAN results to an entire LSAY cohort (rather than a sub-sample).

When a new LSAY cohort commences as part of the PISA assessments (given NAPLAN has been in place since 2008) it will be possible to link LSAY records to NAPLAN data from Years 3, 5, 7 and 9 (see figure 4). A more complete linking of NAPLAN and LSAY data will produce additional achievement data at multiple life stages allowing researchers to determine the impact of academic achievement at Years 3, 5, 7 and 9 on young people's transitions from school to work and their later academic and career outcomes. It would also enable researchers to control for academic achievement at earlier ages and analyse literacy and numeracy development from Years 3 to 9. Questions that examine how well NAPLAN can predict future 'success' and whether those who show the strongest growth across the NAPLAN years have more successful long-term outcomes could also be investigated.

Figure 4 NAPLAN and LSAY linkage options for an LSAY 2015 commencing cohort (for example)



Note: NAPLAN assessments are based on year level while PISA assessments are age-based. This means that PISA participants are 15 years old when they complete the PISA assessment but can span a range of year levels.

It is critical that an expanded dataset containing linked LSAY and NAPLAN data be made accessible for research. To accommodate more detailed and robust analyses the dataset should contain all available NAPLAN variables. Additional data on schooling, collected as part of NAPLAN, could also be used to broaden the data available on overall school performance and resources.

The linking of NAPLAN scores to LSAY would also provide an opportunity for future LSAY cohorts to be rebuilt by using NAPLAN scores in place of PISA scores for new additions to the LSAY sample. This would enable issues of differential attrition to be tackled and for the LSAY sample to continue to be representative of the youth population.

Further developments should also include consideration of linkages with other data sources, such as the National Schools Statistics Collection, the ABS Census of Population and Housing (to obtain data on the areas in which respondents live, attend school or undertake further post-school study) and Medicare data.

References

Borgonovi, F & Guillermo, M 2012, *Parental involvement in selected PISA countries and economies*, OECD education working paper no. 73, viewed July 2015, <http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/WKP(2012)10&docLanguage=En>.

Cohen, J 1988, *Statistical power analysis for the behavioural sciences,* 2nd edn, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Daraganova, G, Edwards, B & Sipthorp, M 2013, *Using national assessment program — literacy and numeracy (NAPLAN) data in the Longitudinal Study of Australian Children (LSAC)*, Australian Institute of Family Studies, Canberra.

Data Linkage WA, 2015, viewed June 2015, <http://www.datalinkage-wa.org.au>.

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

Howieson, C, Croxford, L & Howart, N 2008, *Meeting the needs for longitudinal data on youth transitions in Scotland: an options appraisal*, Scottish Government, Edinburgh.

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.

Karmel, T 2013, *Are we there yet? Overview of the Longitudinal Surveys of Australian Youth,* NCVER, Adelaide.

Nakagawa, S & Schielzeth, H 2012, 'A general and simple method for obtaining R^2 from generalized linear mixed-effect models', *Methods in Ecology and Evolution,* vol.4, no.2, pp.133—42.

National Health and Medical Research Council 2007, *National statement on ethical conduct in human research*, Commonwealth of Australia, Canberra, viewed April 2015 <http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf>.

National Statistical Service 2010, *High level principles for data integration involving commonwealth data for statistical and research purposes*, viewed April 2015, *<http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/7afdd165e21f34fdca2577e400195826/$FILE/High%20Level%20Principles%20for%20Data%20Integration%20Involving%20Commonwealth%20Data%20for%20Statistical%20and%20Research%20Purposes.pdf>.*

——2013a, 'Data linkage information series', information sheets, viewed April 2015, <http://www.nss.gov.au/nss/home.nsf/pages/Data%20integration%20-%20data%20linking%20information%20sheet%20three>.

——2013b, *Data integration involving commonwealth data for statistical and research purposes: risk assessment guidelines*, viewed April 2015, *<http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/59fd060543b4e9e0ca257a4e001eacfe/$FILE/Risk%20Assessment%20Guidelines%20-%20December%202013.pdf>.*

Office of the Australian Information Commissioner 2014, *Australian Privacy Principles guidelines*, viewed March 2014, <http://www.oaic.gov.au/images/documents/privacy/applying-privacy-law/app-guidelines/APP-guidelines-combined-set-v1.pdf>.

——2015, 'The Privacy Act', viewed March 2015, <http://www.oaic.gov.au/privacy/privacy-act/the-privacy-act>.

Population Health Research Network, 2011, viewed June 2015, <http://www.phrn.org.au/about-us/overview>.

Appendix A: Topic areas from the LSAY, PISA and NAPLAN datasets

LSAY respondents are asked a range of questions across a number of topic areas, as outlined in table A1. For further details about the variables collected as part of LSAY, please refer to the section 'The LSAY data' in the *LSAY 2009 cohort user guide* available at <http://www.lsay.edu.au/publications/2547.html>.

Table A1 LSAY topic areas, LSAY 2009 cohort

Major topic area	Sub-major topic area	Minor topic area
Demographics	Student	Place of residence
	Parent	Education Occupation
Education	School	School characteristics Student characteristics Perceptions about self and school Subjects/courses Subjects/courses: VET Work experience Workplace learning (VET) Careers advice School plans Qualifications and results Government payments and income
	School transition	Main activity Post-school plans School leavers
	Post-school	Study Current study Past study Satisfaction with study Deferred/withdrew from study Apprenticeships/traineeships Current apprenticeships/traineeships Past apprenticeships/traineeships Changed/stopped apprenticeship/traineeship Changed/left employer Changed course Changed institutions Careers advice Government payments and income
Employment	Current	Employment characteristics Time worked Wages and benefits Starting work Working in a job post-school Job training Job satisfaction Perceptions about work Looking for work
	Job history and training	Employment characteristics Time worked Wages and benefits Job training Leaving work

Major topic area	Sub-major topic area	Minor topic area
	Seeking employment	Job search activity Looking for work Problems looking for work
	Not in the labour force	Main activity Education Employment
Social	Health, living arrangements and finance	Living arrangements Children Marriage Disability and health Government payments Housing payments Finance
	General attitudes	Life satisfaction Leisure Volunteer Aspirations Job aspirations and expectations

LSAY topic areas have been used to categorise information available from the PISA 2009 student data file (see table A2). Additional information about schools participating in PISA is available from the school data file.

PISA international student and school data files are available from the PISA 2009 database <https://pisa2009.acer.edu.au/>, and LSAY data can be matched to the PISA international data files by filtering for Australian records using country identifiers (CNT, COUNTRY), and student and school identifiers (STIDSTD and SCHOOLID).

For further details about the variables collected as part of PISA, refer to the section 'Programme for International Student Assessment' in the *LSAY 2009 cohort user guide* available at <http://www.lsay.edu.au/publications/2547.html>.

Table A2 PISA topic areas, PISA 2009

Major topic area	Sub-major topic area	Minor topic area
Demographics	Student	Date of birth/age Gender Indigenous status Country of birth Language spoken at home Socioeconomic status
	Parent	Occupation Education Country of birth Socioeconomic status
Education	School	School characteristics Student characteristics Student achievement Perceptions about self and school Time spent learning Reading activities Reading for school Reading tasks Teaching and learning English Use of computers School plans Science career Subjects/courses: VET Work experience Workplace learning (TAFE) Workplace learning (VET) Libraries
	School transition	Post-school plans
Employment	Current	Working in a job while at school Employment characteristics Time worked Wages and benefits
Social	Health, living arrangements and finance	Living arrangements Household possessions
	General attitudes	Leisure Job aspirations and expectations

LSAY topic areas have been used to categorise student and school-level information collected as part of NAPLAN in 2008 (see table A3). Other information available on the NAPLAN datasets includes:

- student background variables (for example, Australian citizenship, permanent resident, date arrived in Australia)
- ACARA school profile (for example, number of full-time and part-time teaching and non-teaching staff, total student enrolments (by gender), percentage of Indigenous enrolments, proportion of students from language backgrounds other than English (LBOTE) who participated in NAPLAN, attendance rates).

Further details about the variables collected as part of NAPLAN are available from ACARA's data catalogue available at <http://www.acara.edu.au/verve/_resources/Data_Catalogue.pdf>.

In addition to the annual literacy and numeracy assessments in Years 3, 5, 7 and 9, NAPLAN also has triennial sample assessments in science literacy (Year 6), information and communication technology literacy (Years 6 and 10), and civics and citizenship (Years 6 and 10).

Table A3 NAPLAN topic areas, NAPLAN 2008

Major topic area	Sub-major topic area	Minor topic area
Demographics	Student	Date of birth Gender Indigenous status Language background (LBOTE) Citizenship status Language spoken at home Country of birth
	Parent	Education (school) Education (non-school) Occupation
Education	School	State/territory School name Year level School ID Student ID Geographic location Sector Test results – reading, writing, language conventions (spelling, grammar and punctuation) and numeracy

Source: ACARA data catalogue <http://www.acara.edu.au/verve/_resources/Data_Catalogue.pdf>.

Appendix B: Risk assessment

Risk framework

The *Data integration involving Commonwealth Data for statistical and research purposes: risk assessment guidelines* (National Statistical Service 2013b) identify the following eight dimensions to assess the risk of a project:

- sensitivity
- size — refers to the number of identifying variables and identifying information about a data provider
- nature of data collection — refers to 'consent' as this is the main component of the data collection being undertaken for linkage
- technical complexity — refers to the challenges of appropriately confidentialising information
- managerial complexity
- duration of project
- how the data is to be linked
- nature of access.

The risk framework focuses on assessing the risk of a breach of confidentiality and privacy. Three dimensions influence the consequence of a breach. They are:

- sensitivity
- nature of data collection (consent)
- size (information about a data provider).

The remaining five dimensions influence the likelihood of a breach. They are:

- technical complexity
- managerial complexity
- duration of project
- how the data is to be linked
- nature of access.

As a guide, the overall likelihood risk is:

- 'high' if three or more dimensions have been assessed as 'high'
- 'low' if no dimensions are rated 'high' and fewer than three are rated medium.

Mitigation strategies should reduce the likelihood risk considerably.

LSAY–NAPLAN data-linkage risk assessment

The risk assessments were undertaken by NCVER in collaboration with the Department of Education and Training (the Commonwealth data custodian) for the LSAY—NAPLAN data-linkage project. It involved the following three main steps:

- *pre-mitigation risk assessment* - following consultation with the state and territory test administration authorities (NAPLAN data custodians)
- *mitigation strategies* - developed in consultation with the fieldwork contractor and a number of NAPLAN data custodians
- *post-mitigation risk rating* **-** to determine whether an accredited integrating authority is required.

Table B1 provides a summary of the assessment for each of the dimensions used to evaluate the risk of the project. Some of the factors which were important considerations as part of the risk assessment can be found in Box B1.

The LSAY-NAPLAN data linkage project was determined to be 'low risk'. The risk assessment process took about three months to complete.

Table B1 LSAY-NAPLAN data-linkage risk assessment summary

Risk assessment	Dimension	Impact
Pre-mitigation risk assessment		
Consequence of a breach	Sensitivity	Medium
	Consent	Low
	Size	Low
	Final assessment	**Low**
Likelihood of a breach	Managerial complexity	Medium
	Nature of access	Low
	Duration of project	Low
	Likelihood of identification	Low
	Technical complexity	Low
	Final assessment	**Low**
Mitigation strategies		
To reduce consequence of a breach	Sensitivity	Low
	Final assessment	**Low**
To reduce likelihood of a breach	Sensitivity	Medium
	Final assessment	**Low**
Overall rating		**Low**

Box B1 Eight dimensions of risk considered as part of the LSAY–NAPLAN data-linkage risk assessment

Sensitivity

- The data do not include highly sensitive information such as religious beliefs or political opinions.
- Current LSAY data collection and reporting protocols prevent NCVER and researchers from accessing personal information such as contact details.

Consent

- Informed consent is obtained from study participants.

Information about a data provider[1]

- The final linked file does not contain any personal identifiers or information that could be used in identifying individuals or schools.

Managerial complexity

- While the number of agencies and processes involved increases the level of complexity of the project, most agencies involved were experienced in these processes.

Nature of access

- The files of consenting respondents from the fieldwork contractor to the jurisdictions are transferred by email in encrypted format.[2]
- NCVER data holdings at unit record level occur via a secure computer server and staff must sign an undertaking regarding appropriate use of data.
- Access to the final linked dataset is granted to approved NCVER staff only.

Duration of the project

- The file of consenting participants will be destroyed by each data custodian on completion of the project.
- The final linked dataset will be retained by NCVER for five years from the date of publication of the research report.

Likelihood of identification

- The final linked dataset will not contain any identifying information.

Technical complexity

- Only aggregate results will be reported and the linked dataset will not be published.

[1] Current data reporting protocols maintain the confidentiality of the schools participating in PISA by assigning random school identifiers to PISA (and LSAY) records. To ensure that confidentiality at the school level is maintained, these reporting protocols also apply to the LSAY and NAPLAN data-linkage records, and data analysis from the resulting linked file cannot identify schools.

[2] This involved zipping the file and adding a password to open the zipped file. The password was provided by the fieldwork contractor to the nominated officer over the phone. This process ensured that all data contacts were known and limited.

Appendix C: Obtaining consent – guidelines and requirements

According to the Australian Privacy Principles guidelines (Office of the Australian Information Commissioner 2014), the four key elements of consent are:

- the individual is adequately informed before giving consent
- the individual gives consent voluntarily
- the consent is current and specific
- the individual has the capacity to understand and communicate their consent.

The *National Statement on Ethical Conduct in Human Research* 'requirement for consent' conditions are that consent should be voluntary and should be based on sufficient information and adequate understanding of both the proposed research and the implications of participation in it. Respondents must have the opportunity to ask questions and to discuss the information and their decision with others if they wish (National Health and Medical Research Council 2007).

Those who elect not to participate in a research project need not give any reason for their decision. Researchers should do what they can to see that people who decline to participate will suffer no disadvantage as a result of their decision.

Consent can be express or implied.

- Express consent is given explicitly, either orally or in writing. This could include a handwritten signature, an oral statement, or use of an electronic medium or voice signature to signify agreement.
- Implied consent arises where consent may reasonably be inferred in the circumstances from the conduct of the individual and the APP entity.[1]

The Privacy Act does not specify an age after which individuals can make their own privacy decisions. The APP guidelines offer the following:

> As a general principle, an individual under the age of 18 has capacity to consent when they have sufficient understanding and maturity to understand what is being proposed.
>
> If it is not practicable or reasonable for an APP entity to assess the capacity of individuals under the age of 18 on a case-by-case basis, the entity may presume that an individual aged 15 or over has capacity to consent, unless there is something to suggest otherwise.
>
> Office of the Australian Information Commissioner 2014, p.11

This means that LSAY respondents who enter the program when they are, on average, 15 years old are presumed to have the capacity to provide consent at any time during their participation in the program.

[1] An 'APP entity' is defined to be an agency or organisation (Office of the Australian Information Commissioner 2014).

Obtaining consent

The following strategies were used to ensure that consent was gathered appropriately and to the highest standard:

- ‘Double opt-in’ approach - consent was gained initially ‘in principle’ and then confirmed after further information provided and any questions answered.
- Respondents providing oral consent via their annual telephone interviews had their responses recorded as evidence of consent.
- Additional information about the project was made available online. Respondents who were contacted for written consent were posted a copy of the information sheet.

State and territory requirements

As each jurisdiction has different requirements for releasing NAPLAN student-level records, it is important to ensure that the method for obtaining consent (that is, written, telephone or online) is verified by each authority prior to seeking that consent. If the method for obtaining consent is not appropriately verified, then the release of student records can be refused (even if consent has been obtained).

A small number of jurisdictions requested that:

- written consent is obtained from respondents[1]
- the test administration authority is named in the consent form/script (making it clear that respondents were providing their consent to have their NAPLAN records released by that authority).[2]

Written consent was subsequently sought from individuals from one of these jurisdictions but resources and timing constraints prevented written consent being sought from any other jurisdiction.

Most jurisdictions were satisfied with oral (recorded via the telephone interview) and/or online consent on the condition that consent was sought appropriately. Online consent was subsequently gathered from the remaining jurisdictions, but due to budget constraints oral consent was sought from one jurisdiction only.

Numbers of respondents asked to provide consent

Not all respondents had an opportunity to provide their consent; this depended on:

- the format of their interview, that is, whether it was completed online or by phone
  - respondents participating in the third stage (online only) did not have an opportunity to provide their consent if they completed their LSAY interview by phone
- the timeframe for completing their interview
  - respondents participating in the third stage (online only) did not have an opportunity to provide their consent if they completed their LSAY interview before the official consent-gathering period (that is, before August 2014)[3]

[1] Revisions to state or territory acts or data release protocols means these requirements may change in future.

[2] As a result of this requirement, the name of the relevant test administration authority was named in the consent form/script for all jurisdictions.

[3] With the exception of respondents from one jurisdiction who were able to provide their consent during the 12-week period from July to October 2014.

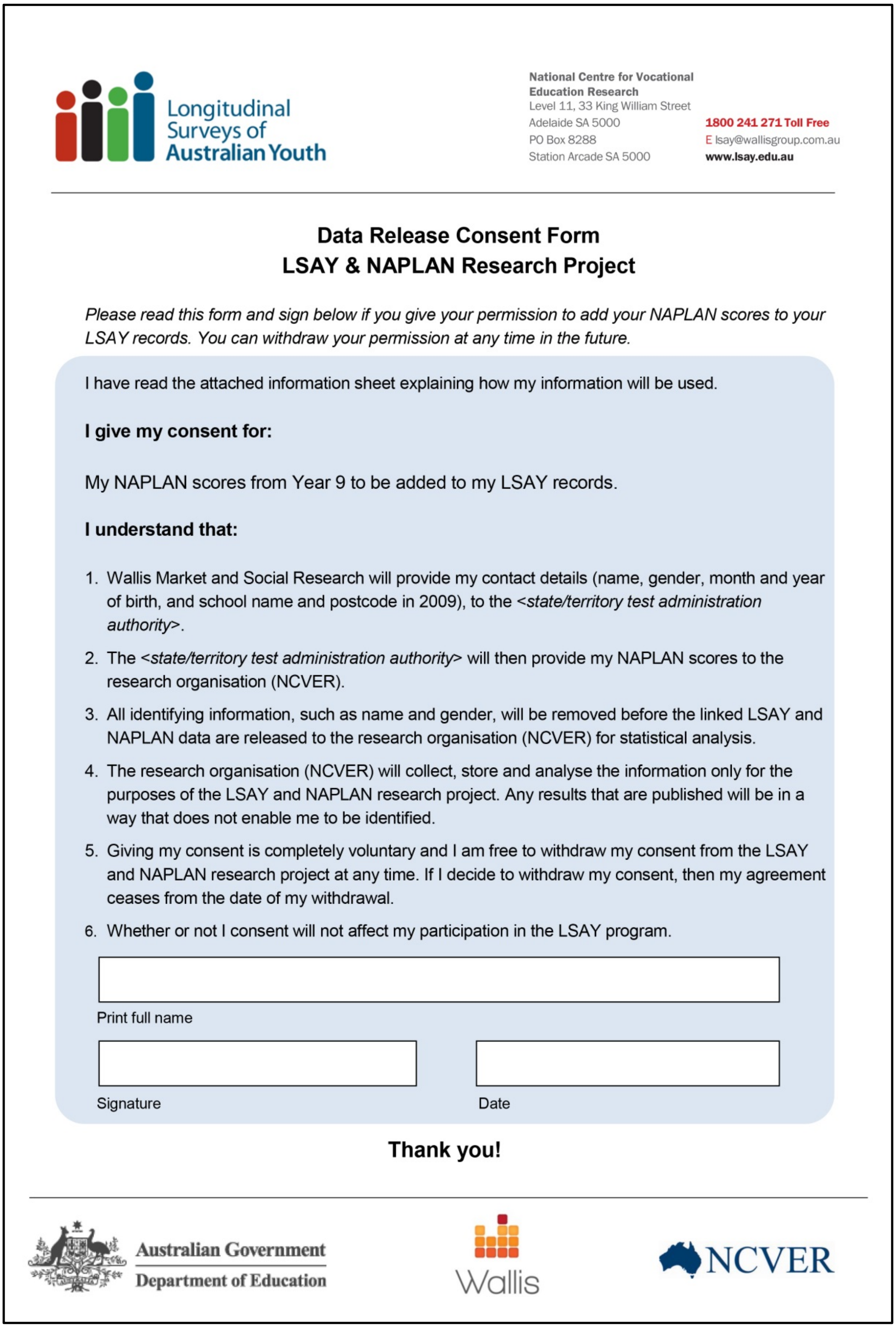
  - respondents participating in the second or third stages did not have an opportunity to provide their consent if they completed their LSAY interview after the consent-gathering period (October 2014).

This means that fewer than half of those eligible to provide consent had an opportunity do so (see table 10). The total number of respondents who were asked to provide their consent and the number of consenting respondents is shown in table 10. Of the 3800 respondents who were eligible to be contacted for consent:

- 1644 were asked to provide their consent, of whom
  - 732 provided their consent
  - 912 did not provide their consent; this includes 725 respondents who did not return their written consent form within the available time period
- 2156 did not have an opportunity to provide consent because they completed their interview by telephone, or because of the timing of their annual interview.

Appendix D: LSAY-NAPLAN data linkage consent form and text

Figure D1 LSAY–NAPLAN data-linkage written consent form



Longitudinal Surveys of Australian Youth

National Centre for Vocational Education Research
Level 11, 33 King William Street
Adelaide SA 5000
PO Box 8288
Station Arcade SA 5000

1800 241 271 Toll Free
E lsay@wallisgroup.com.au
www.lsay.edu.au

Data Release Consent Form
LSAY & NAPLAN Research Project

Please read this form and sign below if you give your permission to add your NAPLAN scores to your LSAY records. You can withdraw your permission at any time in the future.

I have read the attached information sheet explaining how my information will be used.

I give my consent for:

My NAPLAN scores from Year 9 to be added to my LSAY records.

I understand that:

1. Wallis Market and Social Research will provide my contact details (name, gender, month and year of birth, and school name and postcode in 2009), to the <*state/territory test administration authority*>.
2. The <*state/territory test administration authority*> will then provide my NAPLAN scores to the research organisation (NCVER).
3. All identifying information, such as name and gender, will be removed before the linked LSAY and NAPLAN data are released to the research organisation (NCVER) for statistical analysis.
4. The research organisation (NCVER) will collect, store and analyse the information only for the purposes of the LSAY and NAPLAN research project. Any results that are published will be in a way that does not enable me to be identified.
5. Giving my consent is completely voluntary and I am free to withdraw my consent from the LSAY and NAPLAN research project at any time. If I decide to withdraw my consent, then my agreement ceases from the date of my withdrawal.
6. Whether or not I consent will not affect my participation in the LSAY program.

Print full name

Signature

Date

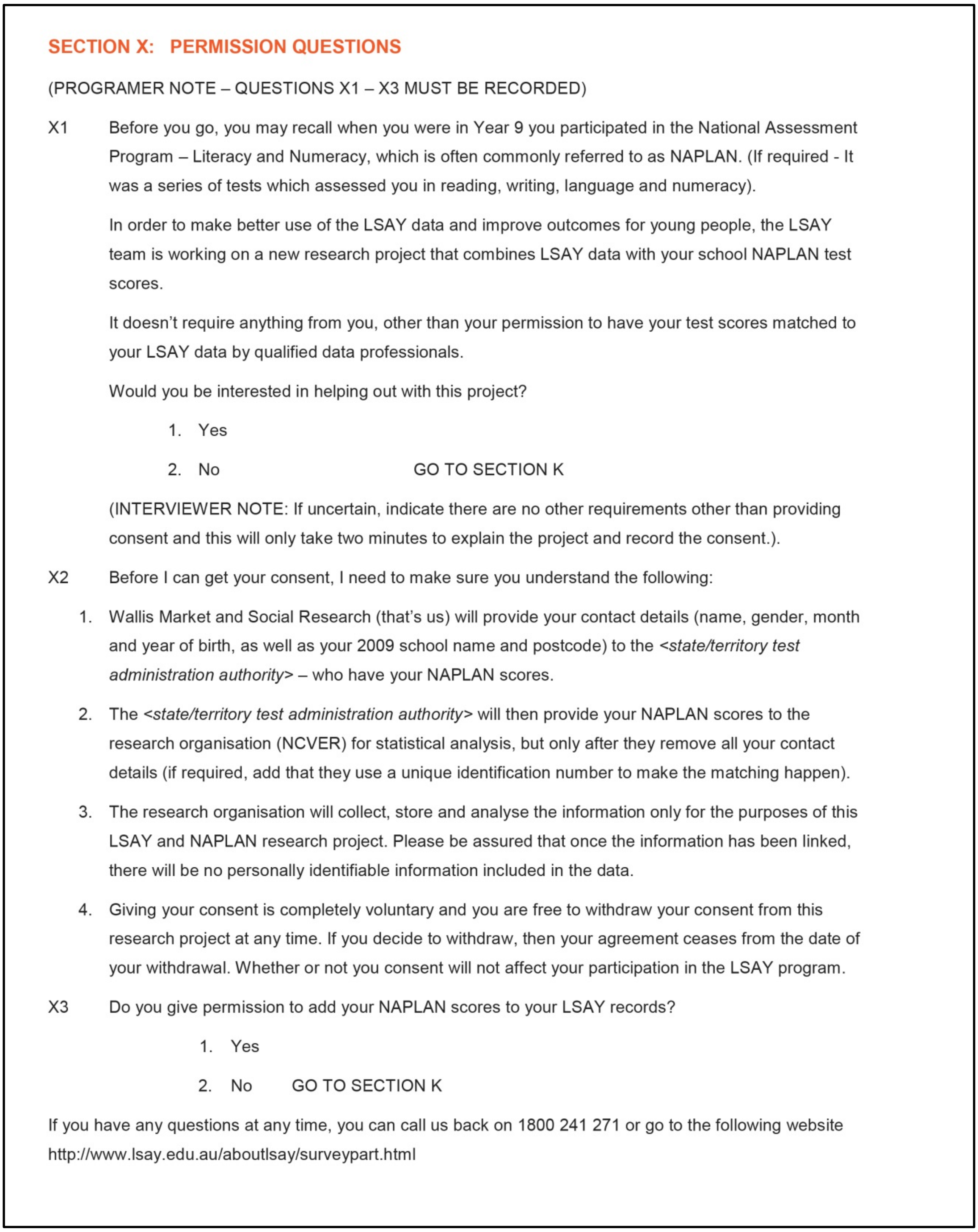
Thank you!

Australian Government
Department of Education

Wallis

NCVER

Figure D2 LSAY–NAPLAN data-linkage telephone and online consent text



SECTION X: PERMISSION QUESTIONS

(PROGRAMER NOTE – QUESTIONS X1 – X3 MUST BE RECORDED)

X1 Before you go, you may recall when you were in Year 9 you participated in the National Assessment Program – Literacy and Numeracy, which is often commonly referred to as NAPLAN. (If required - It was a series of tests which assessed you in reading, writing, language and numeracy).

In order to make better use of the LSAY data and improve outcomes for young people, the LSAY team is working on a new research project that combines LSAY data with your school NAPLAN test scores.

It doesn't require anything from you, other than your permission to have your test scores matched to your LSAY data by qualified data professionals.

Would you be interested in helping out with this project?

1. Yes
2. No GO TO SECTION K

(INTERVIEWER NOTE: If uncertain, indicate there are no other requirements other than providing consent and this will only take two minutes to explain the project and record the consent.).

X2 Before I can get your consent, I need to make sure you understand the following:

1. Wallis Market and Social Research (that's us) will provide your contact details (name, gender, month and year of birth, as well as your 2009 school name and postcode) to the <*state/territory test administration authority*> – who have your NAPLAN scores.
2. The <*state/territory test administration authority*> will then provide your NAPLAN scores to the research organisation (NCVER) for statistical analysis, but only after they remove all your contact details (if required, add that they use a unique identification number to make the matching happen).
3. The research organisation will collect, store and analyse the information only for the purposes of this LSAY and NAPLAN research project. Please be assured that once the information has been linked, there will be no personally identifiable information included in the data.
4. Giving your consent is completely voluntary and you are free to withdraw your consent from this research project at any time. If you decide to withdraw, then your agreement ceases from the date of your withdrawal. Whether or not you consent will not affect your participation in the LSAY program.

X3 Do you give permission to add your NAPLAN scores to your LSAY records?

1. Yes
2. No GO TO SECTION K

If you have any questions at any time, you can call us back on 1800 241 271 or go to the following website http://www.lsay.edu.au/aboutlsay/surveypart.html

Appendix E: Gantt chart of key project phases

Activity	**2013**				**2014**												**2015**			
	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.
Peer reviews of project proposal					■															
Meet and gain advice from appropriate agencies	■	■	■	■	■															
Risk assessment					■	■	■	■												
Work with jurisdictions on approaches and requirements					■	■	■	■	■	■	■	■	■	■						
Develop consent forms and supporting materials					■	■	■	■	■	■										
Stage 1 fieldwork: written consent (S1)									■	■										
Stage 2 fieldwork: telephone/online consent (S2)											■	■	■	■						
Stage 3 fieldwork: online consent (S3)												■	■	■						
Jurisdictions undertake matching of student records										S1				S2 and S3	■					
Statistical analysis and reporting																■	■	■	■	■

Note: Not all jurisdictions were approached to participate in the project at the outset. A decision to approach all jurisdictions some time after the project commenced extended the timeframes for the project phase that sought to meet jurisdictional requirements.

Longitudinal
Surveys of
Australian Youth

Australian Government
Department of Education and Training

NCVER

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
Web www.ncver.edu.au Email ncver@ncver.edu.au