TECHNICAL PAPER

NCVER

Understanding and using the linked LSAY-NAPLAN data: issues and considerations

Emerick Chew
Somayeh Parvazian
Ronnie Semo

National Centre for Vocational Education Research

Publisher's note

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, or state and territory governments or NCVER. Any interpretation of data is the responsibility of the author/project team.

To find other material of interest, search VOCEDplus (the UNESCO/NCVER international database <http://www.voced.edu.au>) using the following keywords: *data analysis; data collecting; data collection; literacy; longitudinal data; longitudinal study; measurement; measuring instrument; numeracy; outcomes; research; students; youth*

This document should be attributed as Chew, E, Parvazian, S & Semo, R 2023, *Understanding and using the linked LSAY-NAPLAN data: issues and considerations,* NCVER, Adelaide.

This work has been produced by NCVER on behalf of the Australian Government and state and territory governments, with funding provided through the Australian Government Department of Education.

COVER IMAGE: GETTY IMAGES

ISBN 978-1-922801-06-7
TD/TNC 150.05

Published by NCVER, ABN 87 007 967 311

Level 5, 60 Light Square, Adelaide SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

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

About the research

Understanding and using the linked LSAY-NAPLAN data: issues and considerations

Emerick Chew, Somayeh Parvazian and Ronnie Semo, NCVER

This paper seeks to identify and address issues that may arise with the use and analysis of the linked Longitudinal Surveys of Australian Youth (LSAY) and the National Assessment Program — Literacy and Numeracy (NAPLAN) data. The study first explores the factors affecting an LSAY respondent's propensity to consent to the data linkage, the aim being to investigate any resultant bias in the linked data. The study also describes two associated procedures: the application of customised weights to deal with the potential bias; and a representativity analysis, which indicates how representative the consenting respondents are of the target population. Finally, the findings from the study are used as the basis of recommendations and guidance for researchers on the use of the linked LSAY-NAPLAN data in analyses.

Key messages

- Consistent with other studies, those less likely to consent to data linkage are: respondents who speak a language other than English at home; those from metropolitan areas; those with a low socioeconomic background; and those who completed the survey online.
- Subconscious factors, such as a respondent's attitudes or personality traits, also influence a respondent's propensity to consent. Respondents who value cooperation more highly are less likely to consent, perhaps because consent does not involve their active and ongoing participation in the data-linkage process or perhaps because respondents do not see the immediate benefits of the data linkage.
- Respondents who tend to value knowledge, as measured by their epistemological beliefs, are more likely to provide their consent for linkage.
- The application of customised weights demonstrates that the weighted linked NAPLAN scores are representative of the original Programme for International Student Assessment (PISA) sample, on which the LSAY sample is based.
- The results from the representativity analysis (R-analysis) indicate that the composition of the consenting respondents is highly representative of the target population from which the LSAY sample is drawn, which implies that the linked data are capable of producing robust estimates for the target population.
- Making the appropriate adjustments through the use of customised linkage weights and/or tailoring the research design means that the linked NAPLAN data can be used in combination with the LSAY data to allow for a unique longitudinal perspective on academic achievements across multiple stages of schooling and the transition of young Australians from school into adulthood.

Simon Walker
Managing Director, NCVER

Contents

Tables

Introduction

The Longitudinal Surveys of Australian Youth (LSAY) follow several cohorts of young Australians as they move through secondary school into further study, work and other destinations. The LSAY datasets allow for analyses into individual changes, transitions and trajectories across the life course of young Australians, providing a valuable evidence base for exploring a range of policy and research questions.

Data linkage

The increasing availability of administrative and assessment data provides new opportunities to combine LSAY data with external sources, with the aim of improving the breadth of information available from the survey. Data linkage can increase the richness and depth of information by linking to data that might be outside the scope of LSAY.

Linking data from longitudinal surveys to administrative records can also provide measures at additional time points (Calderwood & Lessof 2009). Linking to historical data from before the commencement of the survey program (such as primary school test scores) or variables relating to episodes occurring between survey waves (such as school subjects and completion) can be used to supplement or validate data. Furthermore, the addition of data relating to events after the conclusion of the program or after a student leaves the study (such as post-school education and training activity) allows gaps in the data to be filled.

Literacy and numeracy

Literacy and numeracy performance at school is a key indicator of a young person's trajectory into post-school study and work. Since 2003, the LSAY sample has been drawn from the pool of respondents who participated in the Program for International Student Assessment (PISA), facilitating particularly specialised and authoritative studies and analyses of the influence of student achievement on subsequent education and employment outcomes.

Further opportunities have also arisen with the introduction of the National Assessment Program — Literacy and Numeracy (NAPLAN). This assessment provides a national measure of literacy and numeracy performance for Australia's school-aged population in year levels 3, 5, 7 and 9.

Linking LSAY to NAPLAN achievement scores has the potential to provide more detailed, accurate and objective information about young people's academic achievement at several stages of schooling. This powerful linkage also allows for further insights into the relationship between academic ability at multiple life stages and subsequent educational and employment outcomes.

Obtaining consent

To ensure that the handling of respondents' personal information is undertaken in accordance with the Australian *Privacy Act 1988*, consent must be obtained to undertake the linkage. If the characteristics of those who consent are different from those who do not, then the linked data may not be representative of the LSAY sample of students.

The first part of this study explores the factors that influence a respondent's propensity to consent, in order to help us understand any bias that may be present in the linked data. The section that follows discusses the application of customised weights to assist in dealing with this bias. In the third section we

undertake a representativeness analysis to help us to understand the extent to which our consenting respondents are representative of our target population. Finally, we use the findings from our study to provide researchers with guidance on using the linked data for analyses.

Data and methods

Longitudinal Surveys of Australian Youth (LSAY)

LSAY uses large samples of young people, collecting information across the areas of school and post-school education, employment, household characteristics and key demographics. A two-stage stratified sampling design[1] is used. In the first stage, individual schools are sampled and in the second, individual students within those schools are sampled.

Survey participants in the LSAY collection enter the study at around 15 years of age and individuals are contacted once a year until they reach the age of 25. The first group of students participated in the LSAY program in 1995 (Y95 cohort), and subsequent cohorts were recruited in 1998 (Y98 cohort), 2003 (Y03 cohort), 2006 (Y06 cohort), in 2009 (Y09 cohort) and in 2015 (Y15 cohort).

Programme for International Student Assessment

Since 2003, the LSAY sample has been drawn from the pool of respondents who participated in the Program for International Student Assessment. The sampling unit for these PISA-based cohorts (Y03, Y06, Y09 and Y15) is comprised of 15-year-old Australian school students. These PISA-based cohorts span multiple year levels, where Year 10 is the modal year level.[2]

The availability of PISA assessment data as part of the LSAY dataset has allowed for particularly detailed and rich studies, enabling an analysis of the influence of student achievement on subsequent education and employment outcomes.

The PISA sample (which since 2006 is considered the first wave of LSAY) is based on a two-stage stratified sample design, whereby in the first stage schools are stratified according to a range of explicit and implicit strata in the sampling frame. Schools are then selected with probability proportional to size (PPS) within each explicit stratum. In the next stage, a number of 15-year-old students in that school are randomly selected to participate in the test. Indigenous youth are oversampled in an attempt to provide reliable estimates for this subgroup (Marks & Rothman 2003).

National Assessment Program – Literacy and Numeracy (NAPLAN)

NAPLAN is an annual assessment administered to all Australian students in Years 3, 5, 7 and 9. It tests skills in reading, writing, language conventions (that is, spelling, grammar and punctuation) and numeracy. The data from the NAPLAN tests provide schools with information to measure their students' achievements against national minimum standards and against 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. The data resulting from the NAPLAN tests are collected and stored by each jurisdiction's test administration authority, with each having its own data-release policies and protocols. The Australian

1 The explicit stratification for the PISA 2015 sample included state/territory, sector and modal grade. Certainty selections and implicit stratifications included urbanisation, school gender composition, school socioeconomic level and the International Standard Classification of Education (ISCED) level of the school.

2 The sampling unit for the earlier Y95 and Y98 cohorts was Year 9 Australian school students, with students spanning multiple ages. The modal age when Y95 and Y98 respondents were first surveyed was 14 years.

Curriculum, Assessment and Reporting Authority (ACARA) is the independent authority responsible for developing and managing the National Assessment Program.

Linking LSAY to NAPLAN

The NAPLAN tests were first implemented in 2008. Consequently, the Y15 cohort is the only LSAY cohort to have had the opportunity to participate in NAPLAN testing in the primary years. LSAY respondents from the Y15 cohort were asked for their consent to link their NAPLAN results to their LSAY records as part of their wave 2 (2016) survey.

The addition of linked NAPLAN data at multiple stages of schooling further enhances the suite of academic measures available. In particular, the availability of assessment data at multiple time periods allows studies on the influence of academic achievement on post-school trajectories to be undertaken at earlier schooling stages.

Measurement unit: age vs year level by calendar year

LSAY respondents were, on average, 15 years old when they undertook PISA in 2015 and they span multiple year levels. In contrast, NAPLAN data are collected by academic year level, which spans multiple calendar years. As a result, LSAY data are comprised of young people of the same age but of different year levels, while the NAPLAN data are comprised of young people of the same year level but of different ages.

Table 1 shows the distribution of students across year levels when respondents were surveyed in 2015 as part of PISA. The table also demonstrates the corresponding NAPLAN assessment year levels (in bold text) and calendar years to which the LSAY data could be linked. The boxed row indicates the NAPLAN assessment year level and the calendar year in which the majority of the LSAY respondents participated. That is, most respondents undertook their Year 3 NAPLAN assessment in 2008; Year 5 NAPLAN assessment in 2010; Year 7 NAPLAN assessment in 2012; and Year 9 NAPLAN assessment in 2014.

Given that NAPLAN only commenced in 2008 and that around 14% of LSAY respondents are likely to have been in Year 4 or Year 5 in that calendar year, we would expect to have no Year 3 NAPLAN data for these respondents as they were in Year 3 in 2006 and/or 2007.

Table 1 Y15 respondents by year level in 2015 and corresponding NAPLAN assessment year level and calendar year

	Wave 1 (2015)	**NAPLAN assessment year level and calendar year**									
	%	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Year 7	0.01	Prep	Yr1	Yr2	**Yr3**	Yr4	**Yr5**	Yr6	**Yr7**	Yr8	**Yr9**
Year 8	0.12	Yr1	Yr2	**Yr3**	Yr4	**Yr5**	Yr6	**Yr7**	Yr8	**Yr9**	Yr10
Year 9	11.05	Yr2	**Yr3**	Yr4	**Yr5**	Yr6	**Yr7**	Yr8	**Yr9**	Yr10	Yr11
Year 10	74.63	**Yr3**	Yr4	**Yr5**	Yr6	**Yr7**	Yr8	**Yr9**	Yr10	Yr11	Yr12
Year 11	14.14	Yr4	**Yr5**	Yr6	**Yr7**	Yr8	**Yr9**	Yr10	Yr11	Yr12	
Year 12	0.04	**Yr5**	Yr6	**Yr7**	Yr8	**Yr9**	Yr10	Yr11	Yr12		

Notes

1 Boxed text indicates the modal year level at the first LSAY survey wave in 2015 and the corresponding year level and calendar year.

2 Shaded/coloured text indicates a NAPLAN assessment year level.

3 This mapping assumes a linear trajectory through schooling such that students do not repeat or skip a grade; however, it is likely that some students would have sat the NAPLAN test in a different calendar year.

Sample attrition

Similar to all longitudinal surveys, LSAY suffers from sample attrition. For the LSAY cohorts, attrition is greatest between waves 1 and 2, after which year-on-year attrition rates reduce and stabilise (see table 2). The high rate of attrition between waves 1 and 2 is largely due to the incomplete contact details provided at wave one (Moore & Semo 2019). An analysis suggests that attrition is more pronounced for lower-performing students, those with lower socioeconomic status (SES), and Indigenous students (Australian Department of Education 2016).

Respondents were only asked for their consent to the linkage from wave 2, and, given the attrition experienced between waves I and 2, this means that only a subset of the original LSAY sample was asked for their consent to link their NAPLAN scores to their LSAY records.

Table 2 Sample sizes and response rates, LSAY Y15 cohort, waves 1–5

	Wave/year				
	W1/2015	W2/2016	W3/2017	W4/2018	W5/2019
Sample size (n)	14 530	4 704	4 603	4 825	3 721
% of wave 1	100	32.4	31.7	33.2	25.6
% of previous wave		32.4	97.9	104.8	77.1

Note: Includes respondents recruited as part of the top-up activity in 2017 (wave 3). Further information about the top-up activity is available from the LSAY Y15 user guide (NCVER 2022).

Consent and linkage rates

As noted above, as part of their wave 2 (2016) survey, LSAY respondents from the Y15 cohort were asked for their consent to link their NAPLAN results to their LSAY records. Respondents who did not participate at wave 2 (2016) were asked for their consent in subsequent survey waves. Respondents who responded to the consent question are not asked again in subsequent waves. The consent seeks permission from the respondents to link their NAPLAN data across all year levels: Year 3, 5, 7 and 9.

For details regarding the consent and linkage methodology, refer to the LSAY 2015 cohort user guide (see NCVER 2022).

Table 3 shows the overall consent rate for the LSAY-NAPLAN data linkage is 83%. However, records may not always be successfully linked for consenting respondents because the assessment data may not be available, or details required to undertake the linkage may be missing or incorrect. Nevertheless, the LSAY-NAPLAN linkage achieved an overall linkage rate of 95% for those who provided their consent. Table 3 and 4 show the breakdown of the consent and linkage rates respectively by each wave.

Table 3 LSAY-NAPLAN consent rates, Y15 cohort, waves 2–5

	Wave 2 (2016)		Wave 3 (2017)		Wave 4 (2018)		Wave 5 (2019)		Total	
	n	%	n	%	n	%	n	%	n	%
Provided consent	4004	85	476	80	728	77	79	66	5287	83
Did not provide consent	700	15	121	20	215	23	41	34	1077	17
Respondents asked for consent	4704	100	597	100	943	100	120	100	6364	100

Notes:

1 Includes respondents recruited as part of the top-up activity in 2017 (wave 3).

2 Excludes those who provided their consent but were removed from the LSAY dataset because they were ineligible as part of the PISA guidelines.

Table 4 LSAY-NAPLAN linkage rates, Y15 cohort, waves 2–5

	Wave 2 (2016)		Wave 3 (2017)		Wave 4 (2018)		Wave 5 (2019)		Total	
	n	%	n	%	n	%	n	%	n	%
Provided consent	4004		476		728		79		5287	
Of these: NAPLAN records linked	3842	96	448	94	674	93	75	95	5039	95
Respondents asked for consent	4704		597		943		120		6364	
Of these: NAPLAN records linked	3842	82	448	75	674	71	75	63	5039	79

Notes:

1 Includes respondents recruited as part of the top-up activity in 2017 (wave 3).

2 Excludes those who provided their consent but were removed from the LSAY dataset because they were ineligible as part of the PISA guidelines.

When comparing the number of records that were linked for each NAPLAN assessment year level, the results from earlier year levels were more difficult to retrieve (see table 5). The significantly lower rate of linked NAPLAN records for Year 3 can also be attributed to the fact that a number of LSAY respondents would not have participated in NAPLAN when they were in Year 3 because, as shown in table 1, a large proportion of the respondents would have been in Year 4 or 5 when NAPLAN was introduced, in 2008.

Table 5 LSAY-NAPLAN linked data by academic year level

Academic year	Participants with a linked record (n)	% of total NAPLAN records linked
Year 3	3850	76.4
Year 5	4657	92.4
Year 7	4766	94.6
Year 9	4865	96.5
Total (Years 3, 5, 7 or 9)	**5039**	**100**

Understanding consent bias

Bias arises in data linkages when particular demographics or particular characteristics of the respondent are likely to influence their decision to provide their consent. Considering the relatively high linkage rate achieved — of 95% — our main concern lies with the group of *consenting* respondents and not necessarily the *linked* respondents.

Methodology

A logistic regression model is used to study the potential bias in the consenting respondents. The statistical modelling includes 5813 LSAY respondents, where 4864 are consenting respondents and 949 are non-consenting respondents.

The choice of variables used in the regression modelling is based on informed knowledge of the existing literature describing similar research, combined with an understanding of factors that may suggest consent bias, which are available from the LSAY data.

These factors include indices that are derived using the respondent's perspectives on certain values, such as the value of collaboration or appreciation of knowledge, which would, it is hoped, measure the respondent's subconscious attitude towards data linkage to some extent.

All variables have been checked to avoid for confounding issues, while any non-significant variables, those not adding valuable information in the interpretation of the model, are also excluded. The profile of all variables used in the statistical modelling is shown in table 6.

CPSVALUE is a self-reported collaboration and teamwork index measuring the extent to which the respondent values the concept of cooperation. A higher index indicates higher value for cooperation. EPIST is a self-reported index measuring the respondent's epistemological beliefs about science. A higher index indicates stronger belief.

Mode of consent is the mode of survey completion when responding to the consent question. Mixed mode, that is, a combination of both computer-assisted telephone interview (CATI) and computer-assisted web interview (CAWI), is possible, although uncommon. For simplicity, those who have completed the survey in mixed mode have been reclassified to CATI.

The academic achievement scores of the respondents — specifically, reading achievement scores — have been considered for inclusion in the modelling as a measure of the respondent's literacy. However, due to its intercorrelation with multiple variables (sex, socioeconomic status[3] and the epistemological belief index), the reading achievement scores have been excluded from the model.

Considering the two-stage stratified sample design of the PISA sample on which the LSAY sample is based, mixed-effect modelling could prove to be a better approach, taking into account random effect(s) such as the sample stratum, schools, states or sectors.

However, after performing the analysis, the findings indicated otherwise. The random effects of stratum, schools, states or sectors were not significant and do not indicate any variability in consent between those groups. Therefore, this study uses the standard logistic regression model and does not include any random effects.

3 Socioeconomic status is measured using the PISA index of economic, social and cultural status (ESCS).

Table 6 Summary statistics by NAPLAN consent status

	Provided consent				Total	
	No		Yes			
	n	*%*	*n*	*%*	*n*	*%*
Sex						
Female	626	**17.6**	2922	**82.4**	3548	**100**
Male	451	**16.0**	2365	**84.0**	2816	**100**
Language spoken at home						
English spoken at home	799	**15.3**	4436	**84.7**	5235	**100**
Language other than English spoken at home	150	**26.0**	428	**74.1**	578	**100**
Missing	128	**23.2**	423	**76.8**	551	**100**
Indigenous status						
Non-Indigenous	945	**17.0**	4617	**83.0**	5562	**100**
Indigenous	128	**16.2**	664	**83.8**	792	**100**
Missing	4	**40.0**	6	**60.0**	10	**100**
Mode of consent						
Computer-assisted telephone interview (CATI)	250	**10.3**	2187	**89.7**	2437	**100**
Computer-assisted web interview (CAWI)	827	**21.1**	3100	**78.9**	3927	**100**
Geographic location of school (major categories)						
Metropolitan	724	**17.1**	3510	**82.9**	4234	**100**
Provincial	212	**14.7**	1232	**85.3**	1444	**100**
Remote	16	**10.9**	131	**89.1**	147	**100**
Missing	125	**23.2**	414	**76.8**	539	**100**
Socioeconomic status (ESCS)						
Lowest quintile	181	**20.4**	708	**79.6**	889	**100**
Second quintile	166	**16.5**	839	**83.5**	1005	**100**
Third quintile	201	**17.4**	955	**82.6**	1156	**100**
Fourth quintile	217	**16.6**	1090	**83.4**	1307	**100**
Highest quintile	178	**12.3**	1265	**87.7**	1443	**100**
Unknown	134	**23.8**	430	**76.2**	564	**100**
Value cooperation (CPSVALUE)						
Lowest quintile	177	**12.3**	1266	**87.7**	1443	**100**
Second quintile	166	**15.9**	878	**84.1**	1044	**100**
Third quintile	225	**17.3**	1073	**82.7**	1298	**100**
Fourth quintile	162	**19.1**	685	**80.9**	847	**100**
Highest quintile	214	**18.7**	933	**81.3**	1147	**100**
Unknown	133	**22.7**	452	**77.3**	585	**100**

	Provided consent				Total	
	No		Yes			
	n	*%*	*n*	*%*	*n*	*%*
Epistemological beliefs (EPIST)						
Lowest quintile	186	**22.0**	658	**78.0**	844	**100**
Second quintile	238	**16.5**	1206	**83.5**	1444	**100**
Third quintile	114	**19.6**	468	**80.4**	582	**100**
Fourth quintile	196	**15.3**	1087	**84.7**	1283	**100**
Highest quintile	155	**11.4**	1204	**88.6**	1359	**100**
Unknown	188	**22.1**	664	**77.9**	852	**100**
Total	1077	**17.0**	5287	**83.0**	6364	**100**

Note: For more information about how ESCS, CPSVALUE and EPIST are derived and scaled, refer to chapter 16 of the PISA 2015 technical report, which can be accessed at <https://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-16-Procedures-and-Construct-Validation-of-Context-Questionnaire-Data.pdf>.

Table 7 indicates the variables found to be strongly associated with providing consent for linking the NAPLAN data. Variables with a p-value of less than 0.05 are considered to be statistically significant in this study.

Table 7 Type III tests of fixed effects

Effect	Degrees of freedom	F-Value	Pr > F	Stat. sig. at 5% level
Sex	1	1.12	0.2893	
Language spoken at home	1	24.68	<.0001	*
Indigenous status	1	0.1	0.7547	
Mode of consent	1	107.66	<.0001	*
Geographic location of school (major categories)	2	3.4	0.0334	*
Socioeconomic status (ESCS)	5	4.85	0.0002	*
Value cooperation (CPSVALUE)	5	5.87	<.0001	*
Epistemological beliefs (EPIST)	5	8.79	<.0001	*

Table 8 shows the odds-ratio estimates obtained from the logistic regression analysis. An odds ratio greater than one indicates a higher likelihood of consenting for the category of interest (for example, male) when compared with the reference category (for example, female). Conversely, an odds ratio of less than one indicates a lower likelihood of consenting compared with the reference category. An odds ratio equivalent to one implies an equal likelihood of consenting for both categories.

Table 8 Odds-ratio output from logistic regression analysis on likelihood to consent

		Odds-ratio estimate	Pr > |t|	Statistical sig. at 5% level
Sex (reference: Female)	Male	1.082	0.2893	
Language spoken at home (reference: English)	Language other than English spoken at home	0.586	<.0001	*
Indigenous status (reference: Non-Indigenous)	Indigenous	1.035	0.7547	
Mode of consent (reference: CATI)	CAWI	0.426	<.0001	*
Geographic location of school (reference: Metropolitan)	Provincial	1.174	0.0722	
	Remote	1.746	0.0416	*
Socioeconomic status (ESCS) (reference: Lowest quintile)	Second quintile	1.306	0.0293	*
	Third quintile	1.224	0.0882	
	Fourth quintile	1.261	0.0485	*
	Highest quintile	1.699	<.0001	*
	Unknown	0.422	0.0842	
Value cooperation (CPSVALUE) (reference: Lowest quintile)	Second quintile	0.792	0.0497	*
	Third quintile	0.678	0.0005	*
	Fourth quintile	0.579	<.0001	*
	Highest quintile	0.604	<.0001	*
	Unknown	1.07	0.8958	
Epistemological beliefs (EPIST) (reference: Lowest quintile)	Second quintile	1.383	0.0038	*
	Third quintile	1.201	0.1826	
	Fourth quintile	1.596	<.0001	*
	Highest quintile	2.144	<.0001	*
	Unknown	1.171	0.3697	

The results demonstrate that sex and Indigenous status are the only two demographic factors that do not display a significant association with consenting. On the other hand, language spoken at home, mode of consent, geographic location of school, socioeconomic status, self-reported index of value cooperation and epistemological beliefs demonstrate a strong association with the likelihood to consent.

Discussion

Respondents who speak a language other than English at home are less likely to consent to linkage. If English isn't the respondent's primary language, they may face difficulty in understanding the consent question, thus making their consent less likely (Jäckle, Beninger et al. 2021). Studies by Baghal, Knies and Burton (2014) and Carter et al. (2010) also reported lower consent rates among ethnic groups who

speak multiple languages at home. However, it may not always be the case that English is the main barrier for consent for respondents who speak other languages at home. For example, a respondent speaking multiple languages at home may indicate a diversity of cultures in their upbringing, which in some instances could lead to concern about sharing their personal information online.

Findings from this study also suggest that respondents who completed the survey online (CAWI) are less likely to consent compared with those who completed the survey through a telephone interview (CATI). This is consistent with findings by Jäckle, Beninger et al. (2021), who found that respondents were less likely to provide their consent to data linkage when completing their survey online. Because there is a tendency to skim-read text online, respondents are less likely to read all the information presented on the screen, leading to key aspects of the consent request being missed. In a conversational setting such as a telephone or face-to-face interview, respondents have an opportunity to ask the interviewer clarification questions, resulting in a better understanding of the intent of the data linkage.

Concerns about privacy and the security of personal information shared online is another key reason why respondents may be less likely to provide their consent online. With online completion of the interview, no interviewers are available to address respondents' potential concerns about data linkage, unlike CATI.

The social norms associated with the presence of an interviewer may also change the outcome of an interview (Jäckle, Burton et. al. 2021); for example, in a qualitative study conducted by Beninger et al. (2017), participants cited social pressures to conform as a reason why they might be more likely to give consent in a face-to-face interview as opposed to online.

Geographic location of school demonstrates a significant association with the likelihood to consent, with respondents studying in remote schools more likely to consent than those in metropolitan schools. A compilation of literature reviews by Yang, Fricker and Eltinge (2019) found consistent evidence that socio-environmental features play some role in shaping respondents' decision to consent. Various studies have found that respondents living in urban/metropolitan areas are less likely to consent than those living in non-metro areas (Jenkins et al 2006; Dahlhamer & Cox 2007; Baghal, Knies & Burton 2014).

A further factor displaying a positive association with consent are respondents from the second, fourth and highest quintile of socioeconomic status, with these more likely to consent compared with respondents from the lowest quintile. Bandara et al. 2019 and Baghal, Knies and Burton (2014) reported similar findings, in that those from the lower socioeconomic levels are less likely to provide their consent.

Subconscious factors such as respondents' attitudes or personality traits have been key discussion points in various studies on data-linkage consent (Yang, Fricker & Eltinge 2019). Our statistical modelling has identified two such factors, these making a significant contribution to the likelihood of consent, including how much the respondent values cooperation and the strength of their epistemological beliefs.

Interestingly, respondents who identify themselves in the higher quintiles of valuing cooperation are less likely to consent, relative to those in the lowest quintile. A possible explanation could be that respondents who value cooperation see cooperation as an active and ongoing collaborative process, in which the respondent is involved. Providing consent without being involved in the data-linking or data-usage process may not represent a sufficiently strong enough incentive for cooperation. Another possible explanation could be that respondents do not see the immediate benefits of the data linkage so decide not to allow other agencies to access their data.

A higher index of epistemological beliefs about science reflects a stronger belief about knowledge. The results show that, apart from respondents who have identified themselves in the third quintile, those

who fall in all other quintiles are more likely to provide their consent by comparison with those in the lowest quintile. Notably, respondents in the highest quintile have twice the likelihood of consenting to NAPLAN data linkage relative to those in the lowest quintile, suggesting that respondents with a stronger appreciation of knowledge have a better understanding of the data-linkage process — as well as its potential implications — making them more willing to consent.

Using weights to address bias

The regression analysis suggests that a number of factors are associated with the likelihood of respondents consenting to the data linkage, which can lead to bias when using the linked data. One way to address this bias is to create specialised weights for the linked LSAY-NAPLAN data.

Current weighting methodology

The weights created in LSAY attempt to ensure that the sample matches the original population, given both the use of a complex sampling scheme and the effect of attrition. Two sets of weights are created for each respondent: the base weights (first wave weights) and longitudinal weights.

The first wave weights are calculated by the PISA Consortium (OECD 2017). The weights are based on the sampling scheme employed and the probability of selection of a school and an individual. The weights are constructed to ensure that, when applied, the collected sample represents the underlying population of 15-year-olds attending school in 2015.

From wave 2 onwards, we adjust the PISA sample weights to account for the attrition that occurred between the PISA and LSAY survey waves. LSAY computes weights based upon the propensity to respond and to drop out (Lim 2011). This is a straightforward extension of the propensity score theory of Rosenbaum and Rubin (1983) and is incorporated into survey non-response problems.

Historically, the LSAY weight variables have been constructed by first recalculating the sampling weights and the attrition weights separately. These two weights are then multiplied. The wave-on-wave sample and attrition weights are derived using the logistic regression approach.

Note that only cases who have responded to the latest wave will have weights at that wave. In studies requiring the use of an unbalanced dataset (for example, when the variable under investigation has occurred at different waves or for a different length of time across various numbers of waves for different respondents), we need to either build study-specific custom weights or use other methods to account for the survey design and attrition patterns in the data.

Custom weights: required for analysing specific subsets of the survey

Surveys are subject to many different forms of analyses. Some respondents may have provided all the data needed for certain analyses, but they may not have provided all the data required for other analyses. This means that, if only those providing the requisite data for a particular analysis are included in that analysis, differing analyses will be based on different subsets of the sample rather than on a consistent sample.

This raises a complicating factor, in that different sets of weights are needed according to the subset of the sample included in a particular analysis. For example, in the current linkage study, we are interested in any LSAY participant who has ever consented to NAPLAN linkage. This means that there would be people consenting to linkage in various waves, and even though they might have dropped out of the sample in the final wave, their data may still be available. In this instance, using the final wave LSAY weights would not be adequate. A custom set of weights needs to be created to adjust the estimates provided for this specific research question.

The custom-weighting program used in this project calculates its weights by first developing a new temporary list of individuals who had consented to the linkage; this process creates an unbalanced dataset, whereby individuals had participated in a different number of waves for the survey. Each person is then individually assigned the most recent weight that was recorded for them. These new longitudinal weights are then subsequently raked, truncated and calibrated[4] (see also Izrael, Hoaglin & Battaglia 2000, 2004; Izrael, Battaglia & Frankel 2009) to represent the relevant population at wave 1. A similar adjustment can be made for analysing other specific subsets of the data.

Comparing linked LSAY-NAPLAN scores with national NAPLAN scores

For this study we created a custom set of weights to account for the fact that only a subset of NAPLAN data is available through linkage. We create a unique set of weights for each NAPLAN academic year level (Years 3, 5, 7 and 9) and, given that respondents will complete the NAPLAN assessment in different calendar years, we use the modal year as the reference point for the weightings. The modal year is the calendar year in which the highest number of respondents completed their NAPLAN assessment for that academic year level. For example, most respondents undertook their Year 9 NAPLAN assessment in 2014, their Year 7 NAPLAN assessment in 2012 etc (as shown in table 1).

To investigate how the linked LSAY-NAPLAN scores compare with the national NAPLAN scores, we employ the Wilcoxon signed rank test to determine whether the difference between the NAPLAN scores is significant.

The first step in conducting the Wilcoxon signed rank test is to calculate the difference between the scores of the linked data and the national mean score. The score differences are then ranked according to the magnitude and direction of the differences. We then evaluate the sum of the ranks across both positive and negative differences to obtain the Wilcoxon test statistic. We use the test statistic and the appropriate critical value to conclude whether there is sufficient evidence to suggest a significant difference.

Table 9 compares the NAPLAN scores from the linked data with the national NAPLAN data, along with the results from the Wilcoxon signed rank test. A p-value of less than 0.05 is considered to be statistically significant. An asterisk in the corresponding column indicates there is a significant difference between the mean scores.

The results show that the linked NAPLAN scores are statistically higher than the national NAPLAN scores across all domains and year levels. This is not unexpected, nor does it present an issue in using the linked data. This is because the aim of the weighting process is to account for the survey design variables and non-response (or for this study, non-consent). We do not expect the average of the weighted linked scores to reflect the average national scores because the survey design is based on other factors.

For example, as discussed earlier, age is a sampling-design variable for PISA, by comparison with NAPLAN, where year level is used. Some young people may no longer be in school at the age of 15 at the time of the PISA assessment, and this can vary across states and territories, depending on the school starting age. This means some of the lowest achievers in the age cohort may be excluded from PISA but included in NAPLAN.

4 Raking is a widely used technique for developing survey weights. It assigns a weight value to each sampling unit such that the weighted distribution of the sample is in very close agreement with two or more marginal control variables. Weight truncating refers to increasing the value of extremely low weights and decreasing the value of extremely high weight values to reduce their impact on the variance of the estimates, especially for subgroup estimates.

Nevertheless, given that the survey design variables and the factors contributing to non-response (or consent bias) are taken into account in the weighting process, we can be confident that the weighted NAPLAN scores are representative of the original PISA sample.

Table 9 Year level and the corresponding modal year in LSAY-NAPLAN data

Year level and modal year	LSAY-NAPLAN mean (weighted)	National NAPLAN mean	P-value	Statistical sig. at 5% level
Year 9, 2014				
Reading	592.1	580.4	<.0001	*
Writing	564.4	550.3	<.0001	*
Spelling	596.9	582.0	<.0001	*
Numeracy	599.4	587.8	0.0002	*
Grammar & punctuation	586.2	573.5	<.0001	*
Year 7, 2012				
Reading	553.5	541.5	<.0001	*
Writing	528.7	518.3	0.0009	*
Spelling	556.9	543.4	<.0001	*
Numeracy	552.4	538.1	<.0001	*
Grammar & punctuation	558.4	546.2	<.0001	*
Year 5, 2010				
Reading	499.2	487.4	<.0001	*
Narrative writing	497.6	485.2	<.0001	*
Spelling	499.3	487.1	<.0001	*
Numeracy	499.5	488.8	0.0018	*
Grammar & punctuation	512.3	499.7	<.0001	*
Year 3, 2008				
Reading	418.0	400.5	<.0001	*
Narrative writing	426.9	414.2	<.0001	*
Spelling	416.3	399.5	<.0001	*
Numeracy	411.6	396.9	<.0001	*
Grammar & punctuation	421.2	403.2	<.0001	*

Representativeness analysis

Our previous analysis found that several factors affect the likelihood of a respondent providing their consent to linking their NAPLAN scores to their LSAY records (see section 'Understanding consent bias'). While the presence of these factors introduces an element of bias into our linked data, it does not imply that our consenting respondents are not representative of the population from which the sample is drawn. To investigate this further, we undertake a representativity analysis (also known as an R-analysis) to understand how well our consenting respondents represent the target population.

Schouten, Cobben and Bethlehem (2009) developed a set of measures called 'representativity indicators' (or R-indicators), which measure the degree to which survey respondents resemble the complete sample (Parvazian 2022). The R-indicators are the response propensities from a probit or logit model and are calculated using a set of variables from the population that the sample should represent. Like any propensity value, the R-indicator value ranges from zero to one, where a value of one implies full representativeness and zero implies no representativeness.

Methodology

The R-analysis uses a subset of the data (that is, data for consenting respondents) and evaluates its similarity to the full dataset. Because the linkage rate of all consenting respondents is very high, at about 95%, we perform the R-analysis on the consenting respondents rather than on the linked respondents, as this gives us a close approximation of the representativeness of the linked data.

As a first step, we compare the characteristics of the (weighted) LSAY (PISA) sample at wave 1 with the national NAPLAN population. Table 10 compares the socio-demographic distribution of the LSAY Y15 wave 1 sample with the national population of students who undertook NAPLAN at Year 9 in 2014. We choose the national Year 9 NAPLAN population in 2014 as this cohort is most likely to be the same group of students who sat PISA in 2015 (as most respondents would have been in Year 10 when they sat PISA in 2015).

Table 10 Socio-demographic distribution of the LSAY Y15 wave 1 PISA sample vs national NAPLAN data (Year 9, 2014)

	LSAY Y15 wave 1 PISA sample (weighted)	National Year 9 NAPLAN population
	%	%
Sex		
Female	49.6	48.8
Male	50.4	51.2
Indigenous status		
Non-Indigenous	95.8	94.9
Indigenous	4.2	5.1
Language background other than English		
English	86.6	75.5
Language other than English	11.2	21.2

	LSAY Y15 wave 1 PISA sample (weighted)	National Year 9 NAPLAN population
	%	%
Missing/not stated/unknown	2.3	3.3
State		
Australian Capital Territory	1.8	1.7
New South Wales	31.3	31.5
Victoria	25.1	24.0
Queensland	20.6	21.5
South Australia	7.1	7.1
Western Australia	11.0	10.8
Tasmania	2.3	2.3
Northern Territory	0.8	1.0
Sector		
Government	57.7	59.3
Non-government	42.3	40.7
Geolocation		
Metropolitan	73.6	72.9
Provincial	25.0	25.2
Remote/very remote	1.4	1.8
Mother's school-level education		
Completed Year 9 or equivalent or below1	5.2	5.6
Completed Year 10 or equivalent	16.7	18.9
Completed Year 11 or equivalent2	9.2	11.5
Completed Year 12 or equivalent	63.3	50.4
Not stated/unknown	5.6	13.6
Father's school-level education		
Completed Year 9 or equivalent or below1	7.1	5.4
Completed Year 10 or equivalent	20.8	18.5
Completed Year 11 or equivalent2	9.1	9.1
Completed Year 12 or equivalent	55.2	40.9
Not stated/unknown	7.8	26.2

Notes:

1 For LSAY wave 1 PISA sample, 'Completed Year 9 or equivalent or below' includes those who had: completed some secondary school, but not more than Year 9; completed primary school only; or did not complete primary school.

2 For LSAY wave 1 PISA sample 'Completed Year 11 or equivalent' is mapped from 'Completed Year 10 or 11 and then did a TAFE Training Certificate III'.

Sources:

Longitudinal Surveys of Australian Youth, Y15 cohort (wave 1, 2015); ACARA NAPLAN deidentified student-level data, Year 9, 2014.

Overall, most of the socio-demographics of the LSAY PISA sample have a similar distribution to the national NAPLAN data. A few exceptions are language background other than English, mother's school-level education and father's school-level education. The LSAY PISA sample has a higher proportion of respondents with an English language background (86.6% vs 75.5% in national NAPLAN); a higher proportion of respondents whose mother completed Year 12 or an equivalent school-level education (63.3% compared with 50.4%); and likewise, a higher proportion of respondents whose father completed Year 12 or equivalent school-level education (55.2% compared with 40.9%). However, these differences may be affected by the high level of unknown responses in the NAPLAN data, particularly when comparing parental education across the two data sources.

Results

The same set of variables used in the logistic model for the consent bias analysis is used in the R-analysis to calculate the response/consent propensity. This includes sex, language spoken at home, Indigenous status, mode of consent, geolocation, socioeconomic status, value of cooperation and epistemological beliefs. The choice of variables used in the R-analysis is based on our understanding from the existing literature or has emerged as important, based on findings from this study.

Table 11 presents the results from the R-analysis. The R-indicator of 0.9126 indicates that we can be confident that our consenting respondents are highly representative of the PISA sample.

Table 11 Sample size, consenting numbers and R-indicator

PISA sample size	Provided consent	R-indicator
14530	4873	0.9126

Note: Excludes respondents recruited as part of the top-up activity in 2017 (wave 3).

The R-analysis also allows us to identify groups that are under-represented or over-represented (based on the set of variables used in the modelling) using variable-level partial R-indicators. Partial R-indicators should be close to zero if they are well represented and contribute the least to the consenting propensities. Relatively large unconditional and conditional values indicate subgroups that are under-represented[5]. Table 12 shows the results for the variable-level partial R-indicators.

Table 12 Variable-level partial R-indicators

Variables	Unconditional	Conditional
Sex	0.003239	0.000177
Language spoken at home	0.005515	0.000898
Indigenous status	0.004187	0.00013
Mode of consent	0.035534	0.02829
Geolocation	0.001761	0.000657
Socioeconomic status	0.007605	0.001147
Value cooperation	0.006708	0.001184
Epistemological beliefs	0.011374	0.001498

5 For further information on conditional and unconditional R-indicators, see Schouten, Shlomo & Skinner (2011, pp. 236—7).

We note that both unconditional and conditional indicators have small values, a result that is not surprising, given how highly representative our consenting respondents are. By comparing the partial indicators, we see that the mode of consent is the largest contributor towards the variance in the response propensity, even after conditioning for that variable. This is followed by the epistemological beliefs of the respondents. However, after conditioning for the variable, the partial R-indicator drops significantly and mode of consent remains as the main factor in the respondent's propensity to consent, showing consistency with our findings from the consent bias analysis.

Discussion

It is important to note that R-indicators rely on auxiliary information to evaluate representativity (Schouten, Cobben & Bethlehem 2009). For our study, we use the PISA survey design variables, along with the factors found to contribute significantly to a person's propensity to consent.

Our results indicate that, despite having obtained consent for only a third of our original sample, the composition of our consenting respondents is highly representative of the population from which our sample is drawn. This means that the linked data can be used to reliably produce estimates for our target population.

Using the linked data: considerations and recommendations

A number of issues need to be considered when using the linked LSAY-NAPLAN data. This section provides data-users with guidance on using the linked data in analyses. It covers important aspects such as the structure of the data file, the notable differences between the NAPLAN and LSAY data, and dealing with non-consent bias through the application of customised weights.

Additional notes about the linked NAPLAN data are available from the LSAY Y15 user guide, available at: <https://www.lsay.edu.au/publications/search-for-lsay-publications/lsay-2015-cohort-user-guide>.

Data file structure

The LSAY datasets are traditionally prepared using a wide format, whereby each row represents one respondent and the variables collected at each survey wave are appended as columns to the data file. In contrast, the NAPLAN data are prepared using a long format, with multiple records for each respondent.

These differences in the structures of the LSAY and NAPLAN data files (and to help manage access restrictions for the different datasets) result in the linked NAPLAN dataset being stored separately from the main LSAY datafile. Users can merge the two files using the LSAYID available on both data files, but when creating the merged file they will need to carefully consider the suitability of either a wide file or a long file for their analysis.

The linked dataset only contains records for those who have had their data successfully linked. Each respondent with linked data will have multiple rows: one for each of the (five) assessment domains (that is, Reading, Writing, Spelling, Numeracy, Grammar & Punctuation) for each of the NAPLAN academic years (that is, Years 3, 5, 7 and 9).

A complete list of the variables contained in the linked dataset can be found in the 'Data linkage' worksheet of the 'Variable listing and metadata' available at: <https://www.lsay.edu.au/publications/search-for-lsay-publications/2621>.

Data access

LSAY unit record files are deposited with the Australian Data Archive (ADA) at the Australian National University. Access to the data is free via a formal request-and-registration process, managed by ADA.

Information about accessing the LSAY data is available from the 'How to access LSAY data' page on the LSAY website at: <https://www.lsay.edu.au/data/access>.

Using the data

As outlined in the 'Data and methods' section, LSAY selects students who are the same age but are in different year levels, while the NAPLAN data are comprised of students in the same year level but are of different ages.

As such, we recommend that users:

- *Analyse the linked data by academic year level and assessment year*

  Respondents will complete their NAPLAN assessment at varying ages and calendar years. The academic year level and assessment (calendar) year are available on both the LSAY and NAPLAN datasets, whereas age is not available on the NAPLAN data file. As such, it makes sense that the academic year level and/or assessment year are used as the reference point for the analysis.

  Users can refer to table 1, which shows the academic year level and the corresponding modal (assessment) year in which the majority of respondents undertook the NAPLAN assessment.

  In addition, because NAPLAN scores are scaled, any given score across the year levels and assessment year represents the same level of achievement over time. This makes it possible to compare NAPLAN scores across assessment years. More information about the NAPLAN assessment scales is available at: <https://www.nap.edu.au/results-and-reports/how-to-interpret/scales>.

  Finally, students who have repeated a NAPLAN assessment are also represented in the dataset. Analysing by academic year level and assessment year would avoid double-counting these records.

- *Exclude reserved values when analysing NAPLAN scores and bands*

  NAPLAN scores and performance bands include reserved values such as -1 and -9 to indicate non-participation in NAPLAN and missing data respectively. These values can be found in the format programs or value labels in the dataset and should be excluded from any numerical analysis.

- *Remove repeated assessment records when analysing an academic year across all assessment years*

  Table 13 indicates the number of respondents with repeated assessments in each year level; for example, one student in Year 9 repeated their NAPLAN assessment. Data-users therefore need to decide which record (that is, assessment year) to retain for their analysis.

Table 13 Number of participants that repeated a year level

Academic year	Number of participants
Year 3	6
Year 5	8
Year 7	6
Year 9	1

- *Avoid making comparisons between the writing assessments conducted in 2008—10 with those from 2011*

For the first time in 2011, students were required to undertake a persuasive writing task. Prior to this, students were required to write a narrative/story for the writing assessment. In the linked LSAY-NAPLAN dataset, the change in the writing assessment is reflected using different variable names: 'NarrativeWriting' for 2008—10; and 'Writing' from 2011.

Applying weights

For this study we created a custom set of weights to account for the fact that only a subset of NAPLAN data is available through linkage. In longitudinal studies with multiple components and subsamples across

multiple rounds of data collection, a number of possible weights can be created to enable analyses of data within and across rounds. However, it is neither economical nor useful in a practical sense to create weights for every combination of components across every round of data collection. Therefore, most surveys advise that individual researchers will need to decide which type of weight is most suited to a specific research question.

In this section, we discuss a few of the different sets of weights that might be needed, depending on the subset of the sample included in a particular analysis of the linked LSAY-NAPLAN dataset.

Summary statistics

One of the most common studies involving linked data includes the presentation of summary statistics. Each study would have a specific requirement, depending on their study span, in order to accurately calculate summary statistics from multiple years of data. For example, a researcher may want to provide summary statistics for LSAY participants who have undertaken the Year 9 NAPLAN test. This research question would involve students who are undertaking the Year 9 NAPLAN test in various calendar years and, even though they might have dropped out of the LSAY sample in various waves, their data may still be required. In this case, a custom set of weights needs to be created to adjust the estimates provided for this specific research topic.

It should be noted that, even though care has been taken to reduce bias as much as possible through weighting, there is still a likelihood of over-estimating national benchmarks using LSAY linked data due to higher rates of dropout from the survey within the lower academic achievement student groups.

Also note that, in this study, in order to create custom weights, the most recent year's data for that specific individual have been used in instances of repeated NAPLAN assessments for an individual.

Longitudinal or cross-sectional data analysis

In other studies, researchers may be interested in an outcome variable other than the NAPLAN score itself. If analysing a balanced dataset, the most recent longitudinal LSAY weight is used. For example, an area of research interest may be the impact of NAPLAN scores on the employment status of people in the final wave of LSAY (age 25). In this case, the latest-wave LSAY weight is used. Similarly, if another specific wave is being examined, then wave-relevant longitudinal weights can be used for that study.

On the other hand, if the analysis is looking into an unbalanced dataset, then a custom set of weights needs to be created, based on that specific outcome variable. For example, looking into the impact of NAPLAN scores on higher education attendance in later stages of life would require custom weights because this example focuses on an outcome variable that occurs for individual respondents at different points in time, meaning that the outcome is recorded in different survey waves. In this instance, each person should be individually assigned the sample weight of the specific wave in which the relevant information on first attendance in higher education was recorded (or in more general terms, in which the most recent information needed for an analysis was recorded). These new weights must then be truncated and calibrated (Izrael, Battaglia & Frankel 2009).

Conclusion

The linked LSAY-NAPLAN data provide an exceptionally rich source of data on academic achievement, which complement the wealth of information already available from LSAY. Available across multiple learning domains at several time points, these assessment data add another important dimension for research, one not possible prior to the linkage.

However, consent must be obtained to undertake the linkage. If the characteristics of those who consent are different from those who do not, then bias may be present in the linked data.

Our analysis explored this potential bias by analysing the factors that influence a respondent's propensity to consent. Consistent with other studies, we found that respondents less likely to consent to data linkage are those who speak a language other than English at home; those from metropolitan areas; those with a low socioeconomic background; and those who completed their survey online.

We also explored how subconscious factors, such as a respondent's attitude or personality traits, may influence their propensity to consent. We found that those with higher levels of cooperation were less likely to consent when compared with those with lower levels of cooperation, the explanation perhaps being that providing consent without an ongoing involvement in the data-linkage process may not be a strong enough incentive for cooperating. Not seeing the immediate benefits of the data linkage could also discourage respondents from sharing their data with other agencies. On the other hand, respondents who tend to value knowledge (measured by their epistemological beliefs) are more likely to consent.

The second part of the study addresses the bias established in the first section. We achieve this by creating customised weights, which we use to test for differences between the LSAY-NAPLAN scores and the national NAPLAN scores. We find that the linked NAPLAN scores are statistically higher than the national NAPLAN scores, noting, however, that this is not a cause for concern. We do not expect the average of the weighted linked scores to reflect the average national scores because the LSAY survey design is based on other factors. We can be confident that the weighted NAPLAN scores are representative of the original PISA sample because the survey design variables and the factors contributing to non-response (or consent bias) are taken into account in the weighting process.

We also undertook a representativity analysis (also known as an R-analysis) to understand how well our consenting respondents represent the target population. Our results indicate that the composition of our consenting respondents is highly representative of the population from which our sample is drawn. This means that the linked data can be used to reliably produce estimates for our target population.

The final section of this study uses our findings to provide data-users with guidance on using the linked data for research and analysis. It covers important aspects of the linked data such as the structure of the data file, notable differences between the NAPLAN and LSAY data, and dealing with non-consent bias through the application of customised weights.

Despite some of the challenges associated with the use of linked data, this paper has demonstrated that, when making the appropriate adjustments by means of customised weights and/or tailoring the research design, the linked NAPLAN data can be used in combination with the LSAY survey data to provide reliable assessment data across several learning domains. This allows for powerful analyses of how academic achievement across multiple stages of schooling influences a young person's life course as they transition from school into adulthood.

References

Australian Government Department of Education 2016, A review of the Longitudinal Surveys of Australian Youth, Department of Education, Canberra, viewed 25 October 2022, <http://www.lsay.edu.au/publications/2844.html>.

Baghal, T, Knies, G & Burton, J 2014, Linking administrative records to surveys: differences in the correlates to consent decisions, Understanding society working paper series, University of Sussex, Institute for Social and Economic Research, no. 2014 — 09.

Bandara, D, Edwards, Mohal, J & Daraganova, G 2019, Consent to data linkage in a child cohort study, Growing Up in Australia: the Longitudinal Study of Australian Children, Centre for Social Research Methods Working paper no, 7/2019, Australian National University, Canberra.

Beninger, K, Digby, A, Dillon, G & MacGregor, J 2017, How people decide whether to give consent to link their administrative and survey data, Understanding society working paper series, University of Sussex, Institute for Social and Economic Research, no. 2017 — 13.

Calderwood, L and Lessof, C 2009, 'Enhancing longitudinal surveys by linking to administrative data' in (ed) P Lynn Methodology of longitudinal surveys, John Wiley & Sons, New Jersey, pp.55-72.

Carter, K, Shaw, C, Hayward, M & Blakely, T 2010, 'Understanding the determinants of consent for linkage of administrative health data with a longitudinal survey', Kōtuitui: New Zealand Journal of Social Sciences Online, vol.5, no.2, pp.53—60.

Dahlhamer, J & Cox, SC 2007, 'Respondent consent to link survey data with administrative records: results from a split-ballot field test with the 2007 National Health Interview Survey', National Center for Health Statistics, Hyattsville, MD.

Daraganova, G, Edwards, B & Sipthorp, M 2013, Using National Assessment Program — Literacy and Numeracy (NAPLAN) data in the Longitudinal Study of Australian Children (LSAC), LSAC technical paper no.8, Australian Institute of Family Studies, Melbourne.

Izrael, D, Battaglia, MP & Frankel, MR 2009, 'Extreme survey weight adjustment as a component of sample balancing (aka raking)', Proceedings from the thirty-fourth annual SAS users group international conference, vol.720.

Izrael, D, Hoaglin, DC & Battaglia, MP 2000, 'A SAS macro for balancing a weighted sample', in Proceedings of the twenty-fifth annual SAS users group international conference, SAS Institute Inc., pp.9—12.

——2004, 'To rake or not to rake is not the question anymore with the enhanced raking macro', in Proceedings of the twenty-ninth annual SAS users group international conference.

Jäckle, A, Beninger, K, Burton, J & Couper, MP 2021, 'Understanding data linkage consent in longitudinal surveys', Advances in Longitudinal Survey Methodology, pp.122—150.

Jäckle, A, Burton, J, Couper, MP, Crossley, TF & Walzenbach, S 2021, How and why does the mode of data collection affect consent to data linkage? Understanding society working paper series, University of Sussex, Institute for Social and Economic Research, no. 2021 — 04.

Jenkins, S, Cappellari, L, Lynn, P, Jäckle, A & Sala, E 2006, 'Patterns of consent: evidence from a general household survey', Journal of the Royal Statistical Society, vol.169, no.4, pp.701—722.

Lim, P 2011, Weighting the LSAY Programme of International Student Assessment cohorts, NCVER, Adelaide.

Marks, GN and Rothman, S 2003, 'Longitudinal studies of Australian youth', Australian Economic Review, vol.36, no.4, pp.428—434.

Moore, J & Semo, R 2019, 'An introduction to the Longitudinal Surveys of Australian Youth (LSAY)', Longitudinal and Life Course Studies, vol.10, no.1, pp.109—123.

NCVER (National Centre for Vocational Education Research) 2022, Longitudinal Surveys of Australian Youth (LSAY) 2015 cohort user guide, NCVER, Adelaide.

OECD (Organisation for Economic Co-operation and Development) 2017, PISA 2015 technical report, OECD, Paris, <https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf>.

Parvazian, S 2022, 'The new phenomena of diminishing survey response: is the latest Longitudinal Surveys of Australian Youth cohort representative of today's young people?', Unpublished.

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

Schouten, B, Cobben, F & Bethlehem, J 2009, 'Indicators for the representativeness of survey response', Survey Methodology, vol.35, no.1, pp.101—113.

Schouten, B, Shlomo, N & Skinner, CJ, 2011, 'Indicators for monitoring and improving representativeness of response', Journal of Official Statistics, vol.27, no.2, pp. 231—253.

Yang, D, Fricker, S & Eltinge, J 2019, 'Methods for exploratory assessment of consent-to-link in a household survey', Journal of Survey Statistics and Methodology, vol.7, no.1, pp.118—255.

National Centre for Vocational Education Research

Level 5, 60 Light Square, Adelaide, SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

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