Australian Education Research Organisation

Longitudinal literacy and numeracy in Australia (LLANIA) dataset

Technical report

August 2023

The Australian Education Research Organisation (AERO) is Australia's national education evidence body, working to achieve excellence and equity in educational outcomes for all children and young people.

Acknowledgement

AERO acknowledges that this publication was made possible by the joint funding it receives from Commonwealth, state and territory governments.

Acknowledgement of Country

AERO acknowledges the Traditional Custodians of the lands, waterways, skies, islands and sea Country across Australia. We pay our deepest respects to First Nations cultures and Elders past and present. We endeavour to continually value and learn from First Nations knowledges and educational practices.

Authors

Dr Wai Yin Wan, Dr Lisa Williams, Dr Eunro Lee and Dr Lucy Lu.

Copyright

How to cite

Wan, W.-Y., Williams, L., Lee, E., & Lu, L. (2023). *Longitudinal literacy and numeracy in Australia (LLANIA) dataset: Technical report.* Australian Education Research Organisation. https://www.edresearch.edu.au/research/technical-papers/longitudinal-literacy-and-numeracy-australia-llania-dataset-technical-report

Publication details

ISBN 978-1-923066-26-7 (online)
Cover image: iStock.com/hxdbzxy

Contents

Tables

Figures

Summary

For the first time in Australia, a longitudinally linked dataset of students' National Assessment Program – Literacy and Numeracy (NAPLAN) participation and results has been created.

- The final linked dataset comprising data from students who were enrolled in the education system from 2008 to 2021 holds data from 6,270,515 students in total, with 25.4% of them (N = 1,594,261) having fully linked data from Year 3 to Year 9.
- Students in Year 3 in 2009 to 2013 and 2015 (N = 1,654,716) were the cohorts that had the potential to reliably complete all 4 rounds of NAPLAN assessment, with fully matched records formed for 81% to 86% across calendar years.

This technical report describes the processes used to arrive at the linkage, as well as the methods used for quality assurance. The result is the creation of the Australian Education Research Organisation (AERO)'s new Longitudinal Literacy and Numeracy in Australia (LLANIA) dataset.

This work has been assessed as being of negligible risk by an external National Health and Medical Research Council registered research ethics review provider as part of a broader research proposal, the feasibility of which relies on the linked dataset described in this paper.

The potential of the LLANIA dataset is significant, with empirical student learning growth now available to incorporate in analyses, as well as in the formation of evidence supporting Australian education policy and practice.

1. Introduction

1.1. Context

Since its inception, the NAPLAN assessment has presented an untapped opportunity. That is, while the NAPLAN tests have been administered to every student in Australia in Years 3, 5, 7 and 9, in every calendar year (except 2020 due to the COVID-19 pandemic), the data have not been automatically linked longitudinally for each student. Gain data, representing one longitudinal link between a student's results in 2 consecutive NAPLAN testing rounds, have been available, but there has not been any further longitudinal information covering the full population.

Several projects have been initiated internationally to capture longitudinal student data. The National Pupil Database (NPD) in England is perhaps the most powerful example of this. It is a linked longitudinal dataset covering tens of millions of students' progress through schooling and higher education. It includes only students attending schools receiving government funding, so falls short of encompassing the entire population, but despite this, it is a large and powerful data resource. Its uses have spanned from studies of school attainment, to providing evidence about the effectiveness of new initiatives, school quality, and the impact of school starting age (Institute for Fiscal Studies, n.d.). Similarly, the National Assessment of Education Progress (NAEP) in the United States provides a rich data resource covering all years of schooling in 10 learning domain areas. However, it too has the limitation of being run using a sample rather than the full population (National Center for Education Statistics, 2022).

In Australia, the Longitudinal Study of Australian Children (LSAC) and Longitudinal Surveys of Australian Youth (LSAY) are both longitudinal data projects following sampled cohorts through their schooling (Australian Institute of Family Studies, n.d.; National Centre for Vocational Education Research, n.d.). The LSAC followed 2 cohorts, each of which included about 5,000 participants initially, across 20 years of schooling and post-school life. The LSAY included 6 cohorts, each of which commenced with between 10,000 and 15,000 participants, and was designed to span 10 years of data (between the ages of 15 and 25) for each student. While both datasets are linked to NAPLAN results, they face limitations in that they are comprised of relatively small samples in the context of the full population of school-aged Australians – a problem that is further exacerbated by attrition rates.

There is a current focus in Australian research on understanding growth in student learning. A report published in 2016 by the Grattan Institute (Goss et al., 2016) made a case for expressing student educational attainment in terms of learning years or learning months, thus enabling comparisons on this time-based metric of different groups of students and spurring a renewed focus on progress in learning. More recently, the potential held in the direct use of linked longitudinal data has been showcased by Larsen et al. (2022), who have examined growth in student learning with a sample of longitudinally linked NAPLAN data in New South Wales and Victoria using latent growth curve analyses.

It is clear that there is significant potential in Australia for the use of a large-scale longitudinally linked dataset on educational attainment. An Australian dataset of this sort would likely play an important role in producing evidence for policy initiatives, designing student- and school-level interventions, and evaluating programs in terms of student learning growth. In addition, enabling the accurate quantification of student learning growth over time opens possibilities for a better understanding of the process of learning itself. All of these avenues speak to the overall enrichment of student and school outcomes in Australia.

The project described in this report is the first to realise the potential of the NAPLAN assessments in linking national student data longitudinally across up to 4 testing rounds, spanning from Year 3 to Year 9. This linked data is the first Australian longitudinal data project to incorporate the whole Australian student population.

This technical report outlines the data linking strategies and decision-making approaches that were used in the production of the Longitudinal Literacy and Numeracy in Australia (LLANIA) dataset. It describes the matching strategies and linkage results for the 13 years of NAPLAN data from 2008 to 2021 and related student details and school data. The linkage sources of the de-identified student-level NAPLAN data and gain data were in long format, where each record had test outcomes for each student in each calendar year, with the number of records in both datasets adding up to over 24 million in total. The linkage project generated the linked data available in both wide and long format for different modelling purposes. In the wide data, each record is populated with NAPLAN participation and outcomes, together with other student and school details linked for all available NAPLAN rounds[1] (Year 3 to Year 9) for each enrolled student through 2008 to 2021. In the long data, an enrolled student has at least one –potentially multiple – records, with each record containing NAPLAN and other details from one NAPLAN round.

1.2. Project aims

The primary aim of this data linkage project was to establish a national longitudinal dataset of students' NAPLAN participation and performance across Years 3, 5, 7 and 9. To achieve this, 2 supporting aims were identified:

1. To use an optimal selection of matching strategies to achieve the maximum number of matched records.
2. To establish the validity of the linkage through rigorous quality analysis.

1 One previous and most recent record is captured for each NAPLAN round in the wide format data if a student sat the tests more than once.

2. Methodology

2.1. Sources of data

The data sources for the linkage were de-identified[2] Stage 2 NAPLAN data and student gain data. The data had been previously collected through standard operations of the Australian Curriculum, Assessment and Reporting Authority (ACARA). Approval for the Australian Education Research Organisation (AERO) to use the data was granted by ACARA, which is delegated by Education Ministers to review and manage requests for access to NAPLAN data. ACARA's Privacy Policy (ACARA, 2022) details its authority to share data for research purposes. In complement, a standing agreement between ACARA and AERO includes use of the data for this project. The key variables drawn from each data source are shown in Table 1.

The Stage 2 NAPLAN dataset (2008 to 2021) used for the linkage project contained around 16 million (16,238,153) records. After removing duplicate records in 2008, 2019 and 2021, about 15 million (14,953,335) records remained. The 66 variables (see Table 1) pertain to administrative and demographic data, and NAPLAN participation and results, respectively. NAPLAN results in this dataset incorporate raw item responses together with the weighted likelihood estimates (WLEs)[3] that are the published final NAPLAN scores for each student. In addition, the dataset holds a set of 5 plausible values (PVs)[4] for each domain, which are proficiency estimates for every enrolled student, including those who were absent or withdrawn from NAPLAN tests. These PVs are used to calculate population statistics such as the population means which are published in the NAPLAN National Report.

The student gain data (2010 to 2021) contained around 10 million (9,821,808) records. After removing duplicate records in 2021, about 9 million (8,930,387) records remained. The 50 variables again covered aspects of administrative and demographic information, as well as NAPLAN results. However, the NAPLAN variables contained 2 successive NAPLAN test points for each student (e.g., Year 3 and Year 5).

Table 1: Data linkage sources and variables

Category	Variables	
Stage 2 Data (N = 14,953,335; Each record was for every enrolled student in each calendar year in Years 3, 5, 7 and 9, from 2008 to 2021)		
Admin & student demography (k = 19)	• calendar year • jurisdiction • year level • school ID • student ID • sector • remoteness	• date of birth (DOB) • gender • Indigeneity • language background other than English • parental education and occupation

2 The datasets received from ACARA do not include student names or addresses. The exact set of variables received from ACARA are listed in Table 1.
3 See the NAPLAN Technical Reports for how WLEs are generated through a psychometric process by ACARA.
4 For details on how PVs are generated, please see the NAPLAN Technical Reports.

Category	Variables	
NAPLAN data for 5 domains (k = 47)	• WLEs • 5 PVs per domain • item responses (for some tests)	• participation status (present, absent, withdrawn, exempt[5]) • test mode (paper, online)
Student gain data (N = 8,930,387; Each record was for every enrolled student in each calendar year with their previous NAPLAN outcomes from 2010 to 2021)		
Admin and student demography (k = 15)	The same admin and demographic variables as in Stage 2 data except for: • year level • remoteness • parental education and occupation	• year level range (e.g., Years 3 and 5, Years 5 and 7) • previous school and student ID for prior test years • same school indicator
NAPLAN data for 5 domains (k = 35)	• previous and current participation • previous and current WLEs	• previous and current test mode • gain scores
School-level data: school attendance rates and levels (N = 108,989, k = 19) from 2014 to 2021 and school profile data (N = 130,379, k = 35); Each record was for every school in each calendar year		
School characteristics	• school IDs • sector • school type (primary, secondary, combined, special) • governing body • year range	• geolocation • Index of Community Socio-Educational Advantage • percentages in Socio-Educational Advantage quarters
School statistics	• attendance rates • attendance levels • enrolment numbers • staffing numbers	• Indigenous and language background other than English (LBOTE) subgroup proportions

Note: N = Number of records; k = Number of variables.

2.2. Matching strategies

The data linkage used 20 matching strategies in total (details are provided in the Results section; see Table 2) that were evolved through the work process. Through each strategy, matched records in both Stage 2 and student gain data were grouped and assigned with a group number. On the completed implementation of the strategies, all matched records and unmatched records from the Stage 2 data were combined to form the longitudinal dataset.

5 There was a small proportion of students who were identified as non-attempt (M), sanctioned (S) or refused (R). A tiny proportion of students had participation status code like I, N, Q, X and Y.

The first 4 strategies were based on student IDs, school IDs and test administration authority (TAA) and their matched previous IDs from the gain data (strategies 1 to 4). The next 8 strategies (strategies 5 to 12) replicated the first 4 strategies and removed presumably redundant leading zeros in some IDs to improve the matching. Later strategies (13 to 16) matched the Stage 2 records that have the same sets of 5 current and previous NAPLAN WLE scores, year level, date of birth (DOB), and gender in the student gain data. In addition, through addressing the yearly differences in the rounding method with different decimal places in the NAPLAN scores, more matched records were achieved. The final group of strategies (17 to 20) went on to utilise date of birth (DOB), TAA and current and previous student IDs.

2.3. Preparation and de-identification of the linked dataset and national sample

School-level datasets of school profiles and student attendance data (refer to Table 1) were all matched for each student and each respective scholastic year to compile the final linked data. Even when there were no matching records at all, single standalone records were maintained in the linked dataset (as listed as strategy 0 in Table 2), ensuring no data loss occurred.

Once the matching exercise was complete, all student- and school- IDs were stripped from the file and every record in the file was provided with a newly generated record ID, student ID, and school ID. These random IDs were generated through a hashing algorithm developed specifically by AERO and cannot be reverse engineered to reveal original values. They were attached to every record in the linked datasets.

Further de-identification strategies were developed to protect against the potential for re-identification of school- and student-level records. For example, AERO ran algorithms to identify schools by counting the number of schools falling into each combination of State, Sector, School Type and Geolocation (e.g., number of NSW Catholic Secondary Remote schools). If any combination of the 4 variables yielded only one school, the school was defined as being immediately identifiable, and if any combination of variables yielded more than one but fewer than or equal to 5 schools, these schools were deemed to be potentially identifiable. Immediately identifiable and potentially identifiable schools were de-identified by suppressing State and Geolocation (i.e., replaced with NA). After suppression, no school fell uniquely into any combination of State, Sector, School Type and Geolocation in any calendar year. Other algorithms testing different combinations of variables were implemented and strategies to de-identify schools were undertaken in a similar way. Secondly, due to the granular nature of Index of Community Socio-Educational Advantage (ICSEA) data, the raw ICSEA values were not included in the final linked dataset as they were considered as having the potential to be used in combination with other variables to identify schools. A newly generated AERO ICSEA value was provided as an approximation of each school's ICSEA value, created to reflect the distribution of the raw ICSEA data across schools. Schools' ICSEA were categorised into 22 bins, ranging from a bin of 0 (low ICSEA) to 21 (high ICSEA). The median ICSEA for the bin an individual school's ICSEA falls in is the AERO ICSEA value provided in the dataset for the school in question.

The raw attendance rate values were not included in the final linked dataset as they were judged to be potentially usable in combination with other variables to identify schools. Attendance rates of each school were classified into 5 groups: attendance rate ranging between 0–60%, 61–70%, 71–80%, 81–90% and 91–100%. The attendance range corresponding to each school is provided in the dataset.

Thirdly, to strengthen the protection of personal and sensitive information, the Date of Birth variable was masked and only Month and Year of Birth were presented in the dataset. Values other than males and females in the Sex variable were randomly distributed across the dataset. Values for the Aboriginal and Torres Strait Islander variable were presented in the dataset at a higher level (recoded to either Y, N, or Unknown/Not Provided from more granular classifications). The final step in data preparation was drawing a national sample from all data in the final linked dataset pertaining to the 4 consecutive Year 3 cohorts (2011 to 2013 and 2015 Year 3 cohorts) that were able to be reliably tracked all the way to Year 9. The sample was drawn to facilitate use of the longitudinal data in research applications, as the full data containing millions of records would not be easily analysable. An additional benefit of drawing the sample was in the further protection this provided against potential re-identification of records. The sample was drawn using a random sampling with proportion allocation technique due to uneven matching rates across jurisdictions. This resulted in a national sample of 697,459 Year 3 students with at least 2 matched records from Year 5 to Year 9. This national sample was tested to be representative of the underlying Year 3 Australian student populations for the period from 2011 to 2015, excluding 2014. Based on the results from the thorough testing conducted, AERO is satisfied that this national sample, which was developed for research purposes, is fully de-identified.

All data processing and linking activities occurred in a highly secure Microsoft Azure environment. All data held in this environment is located in Microsoft Data Centres housed in Australia with access strictly limited. A secure VPN is required for connection to this environment, utilising secure configuration certificates and end-to-end-encryption for data. The environment is partitioned into 3 separate schemas, each with different levels of data classification, aggregation and security controls. Raw data is stored in a raw schema, while transformation, data processing and linking occurred in a staging schema. Both schemas were accessible only to administrative users with strict access controls. After data linkage, the final de-identified full LLANIA dataset (with randomly generated school IDs and student IDs, and with all masking and suppression strategies applied) was loaded to the third secure schema, which is accessible only to administrative users with strict control access. The national sample of the longitudinal data is stored in a Microsoft Teams site accessible only to authorised researchers for approved research projects.

2.4. Quality assurance analysis

The quality assurance (QA) work supporting the data linkage was conducted in 3 stages. The first stage involved AERO personnel external from the linkage team conducting a review, which consisted of 7 tasks to compare group means, standard deviations, and frequencies between the linked data and annual national NAPLAN reports by ACARA. The second stage was an internal review to check the variable value ranges, non-missing and valid newly generated student IDs and school IDs, and missing or non-missing WLEs and PVs by participation status. The third and final stage involved examining the representativeness of the Year 3 students who are fully matched from Year 3 to Year 9 to all Year 3 students in the 6 fully linkable NAPLAN year (2009 to 2013, 2015) cohorts.

Note that this final step of the QA was conducted prior to the application of data suppression and masking strategies, so as to be able to identify as accurately as possible the degree to which the fully matched cohort was representative of the population.

3. Results

3.1. Matching rates by matching strategy

Prior to matching, an initial check of the 2 data sources was conducted. In the student gain data, 31,675 records had blank student IDs and were removed. Subsequently, the total number of records from the 2 data sources reduced from 23,882,722 to 23,852,047. Across all jurisdictions, 91.1% (N = 21,731,804) of the total 23,852,047 records were matched to one or more records within and across the 2 data sources based on 20 matching strategies as shown in Table 2.

Although over 2 million records (8.9%) did not have any matching (see Table 2), these included structural single records such as Year 9 cohorts in 2008 and Year 3 cohorts in 2021, as well as new arrivals in Year 9 every year and any departures from each TAA after being enrolled in only one NAPLAN year.

Around half of the matching (52.3%) was achieved with the first strategy using the student ID, school ID and TAA. The fact that this rate of matching was relatively low reflects the inconsistent student IDs and school IDs across calendar years and TAAs over the Stage 2 and student gain data. The next most efficient strategy was using previous school and student IDs, as well as TAA (14.9%) from the student gain data.

Re-using strategies 1 to 4 and removing leading zeros in student IDs and/or school IDs that were deemed redundant matched 12.6% of the total records. Using the set of 5 domain WLE scores with demographic variables (matching 9.8% of the total records) was also effective.

There was substantial variation in the matching rates by matching strategies across jurisdictions. The variation was primarily contributed by the different degrees of inconsistency in the student IDs and/or school IDs in the NAPLAN database from 2008 to 2021 across jurisdictions. The second contributing factor was student mobility – that is, the rate at which students moved between schools within each TAA and between TAAs. A few jurisdictions had lower matching rates based on the first 12 matching strategies, which only utilised information from student IDs, school IDs and jurisdictions. The additional strategies utilising current and previous NAPLAN WLE scores across 5 domains improved the matching rates to a larger extent for these jurisdictions.

Table 2: Matching strategies and linkage results

Strategy number	Matching strategy	Match count	%
0	No matching record found	2,120,243	8.9%
1	Student ID, ACARA_SML_ID, TAA	12,468,860	52.3%
2	Student ID, TAA_SCHOOL_ID/JURISDICTION School ID, TAA	69,409	0.3%
3	Previous Student ID, Previous ACARA_SML_ID, TAA	170,347	0.7%
4	Previous Student ID, Previous TAA_SCHOOL_ID/ JURISDICTION School ID, TAA	3,555,791	14.9%
6	Student ID, TAA_SCHOOL_ID /JURISDICTION School ID, TAA (leading zeros in the student and school IDs are removed)	11,759	0.05%

Strategy number	Matching strategy	Match count	%
7	Previous Student ID, Previous ACARA_SML_ID, TAA (leading zeros in the student and school IDs are removed)	83,405	0.4%
8	Previous Student ID, Previous TAA_SCHOOL_ID / JURISDICTION School ID, TAA (leading zeros in the student and school IDs are removed)	672,761	2.8%
9	Student ID, TAA_SCHOOL_ID /JURISDICTION School ID, TAA (leading zeros in the student ID removed)	40	0.00%
10	Student ID, TAA_SCHOOL_ID /JURISDICTION School ID, TAA (leading zeros in the school ID removed)	4,406	0.02%
11	Previous Student ID, Previous TAA_SCHOOL_ID / JURISDICTION School ID, TAA (leading zeros in the student ID removed)	158,886	0.7%
12	Previous Student ID, Previous TAA_SCHOOL_ID / JURISDICTION School ID, TAA (leading zeros in the school ID removed)	108	0.00%
13	YLevel, DOB, Sex, Reading_Score, Writing_Score, Spelling_ Score, Grampunct_Score, Numeracy_Score (current and previous scores with one decimal place)	1,544,118	6.5%
14	YLevel, DOB, Sex, Reading_Score, Writing_Score, Spelling_ Score, Grampunct_Score, Numeracy_Score (current and previous scores (whole number)	270,389	1.1%
15	Current Year, YLevel, DOB, Sex, Reading_Score, Writing_ Score, Spelling_Score, Grampunct_Score, Numeracy_Score (using one decimal place)	513,825	2.2%
16	Current Year, YLevel, DOB, Sex, Reading_Score, Writing_ Score, Spelling_Score, Grampunct_Score, Numeracy_Score (whole number))	6,871	0.03%
17	Student ID, DOB, TAA	96,664	0.4%
18	Previous Student ID, DOB, TAA	5,640	0.02%
19	Student ID, DOB, TAA (leading zeros in the student ID are removed)	14,453	0.06%
20	Previous Student ID, DOB, TAA (leading zeros in the student ID are removed)	3,368	0.01%
	Total	23,852,047	100.00%

Note: ID = identifier, ACARA_SML_ID = ACARA simulation school ID, TAA = test administration authority, YLevel = year level, DOB = date of birth, Grampunct = grammar and punctuation.

3.2. Checks for the quality of the linkage and resulting dataset

As described in Section 2.4, the first stage of QA analysis consisted of 7 tasks to compare the linked data with the annual NAPLAN national results published by ACARA. Specifically, for each test year and test cohort, we checked:

- means
- standard deviations (SD)
- counts of students (for all, and by participation categories)
- proportions in achievement bands
- test result ranges.

The main aim of this exercise was to check data completeness and the integrity of the new dataset, ensuring no data loss or corruption through the data linkage process.

The results of the first piece of QA analysis are shown in Table 3. As illustrated, the statistics were consistent with only negligible discrepancies. Since TAA could submit changes to individual students after Stage 2 census data was submitted to ACARA, these small discrepancies are expected. The fact that the relevant counts, means and proportions across bands mirrored very closely the figures contained in the National Reports indicates that the data is complete and of high quality.

Table 3: Stage 1 quality assurance tasks and results

	Tasks	Results
1	Duplicate records in the data per calendar year	No duplicates were detected.
2	Number of students who participated (Participation status – Present)	All counts of participated students matched with the National Report except for 2010 and 2018. In 2010, there were 7 fewer Year 3, 763 fewer Year 7 and 1,102 fewer Year 9 students in the linked data set for Language Conventions and 20 fewer Year 7 students for Reading. These students were specified with a participation status 'Q'. In 2018, 8 to 39 fewer students appeared in the 'Present' category across 5 NAPLAN domains because they were specified with a participation status 'R' – Refused.
3	Number of students in the 3 non-participation categories (Exempt, Absent and Withdrawn) from 2011 onwards	Number of withdrawn and exempt matched up. There were up to 127 fewer students in the absence number in the linked data from 2011 to 2021. These students were specified with a participation status 'S' – Sanctioned.
4	Average of 5 sets of plausible values – mean and SD	National mean and SD matched with those in the National Report for all domains and all year levels.
5	Percentage of results across 6 achievement bands	Band percentages matched with those in the National Report for all combinations of domains, year levels and test calendar years.

	Tasks	Results
6	Maximum and minimum values of the WLE scores in paper-based tests	Maximum and minimum values of WLE scores all matched with the information in the score equivalence tables in the NAPLAN Technical Reports.
7	Number of distinct newly generated school IDs (ACARA_SML_IDs) in Stage 2 data and in linked long data	Linked long data contained one more school than Stage 2 data, possibly contributed by blank School IDs in Stage 2 data.

Note: Each task included checking the comparison to the ACARA national reports by NAPLAN domains and Year levels, and within each calendar year and jurisdiction.

The second piece of QA work was to check:

- variable value ranges
- missing or non-missing WLEs
- PVs by participation status.

Results confirmed the linked data was valid in terms of the checking points.

The final stage of the QA was to examine the representativeness of the Year 3 students who were fully matched from Year 3 to Year 9 to all Year 3 students in the 6 fully linkable NAPLAN year (2009 to 2013, 2015) cohorts. The distributions of different student-level and school-level demographic characteristics were investigated, including:

- parental education
- parental occupation
- gender
- LBOTE
- First Nations
- remoteness
- school sector.

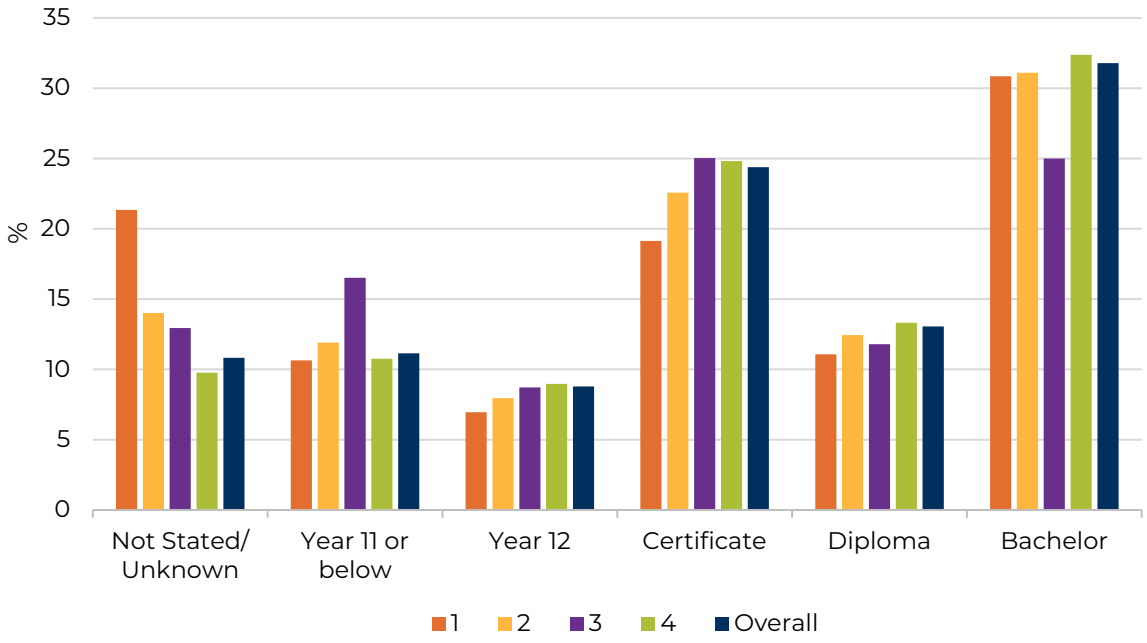
The distributions of all demographics for the fully matched Year 3 students were broadly similar to those of the relevant Year 3 cohorts, with the former group being slightly more socio-educationally advantaged than the latter. Results are shown in Figure 1 for parental education, and indicate that the proportions with different levels of parental education were broadly close, particularly in the comparison between the fully matched cohort (covering 4 NAPLAN testing rounds; green bar) and the overall cohort (navy blue bar).[6] For other demographics and initial Year 3 reading performance, distributions by number of linked records are summarised in Figures A1 to A7 in Appendix A.

6 Of the fully matched Year 3 students, 98% had test results from all 4 test rounds. The vast majority of students who were matched to Year 9 but did not have test results from all 4 test rounds were students who were exempted from one or more of the tests. The Year 3 students who were fully matched to Year 9 and who also had test results from all 4 test rounds were very similar in terms of key demographics when compared to the fully matched Year 3 students.

Compared to the fully matched Year 3 cohorts and the overall cohort, the Year 3 cohorts with 1 to 3 linked records[7] were more likely to be First Nations students and to have their parents' education and occupation not stated or unknown. They were also more likely to reside outside of major cities, and to be enrolled in an Independent school when they were in Year 3.

In terms of Year 3 reading and numeracy performance, the fully matched cohort had slightly better Year 3 performance than the overall cohort, with an average difference of about 3 scale score points for both domains.

Figure 1: Comparisons of parental education across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015



3.3. Linked data

After implementing all 20 matching strategies, 6,270,515 unique students were identified with at least one NAPLAN record through Year 3 to Year 9. Of all unique students, around a quarter (1,594,261) of students had 4 fully linked NAPLAN records from Year 3 to Year 9. A small group of students who had repeated a year of schooling had more than 4 linked records. Nearly one-fifth (1,207,824) of students had 3 linked NAPLAN records in Years 3, 5, 7 or 9. Another 23.1% (1,446,213) of students had 2 linked records. Nearly one-third (2,022,217) of students only had one NAPLAN record, which made up the largest proportion. The primary reason for this was that structurally no linkage was expected for cohorts such as Year 9 in 2008 or Year 3 in 2021. Other possible reasons explaining why some students' results could not be linked include interstate or overseas movements, data unavailability and inconsistency in student IDs over time within jurisdictions. The overall proportion of students having full linkage of records from Year 3 to Year 9 is presented in Figure 2 with further breakdown by jurisdictions.

7 The Year 3 cohorts with 1 to 3 linked records were much fewer in number than those fully matched to Year 9. See relevant statistics in 3.3.

Figure 2: Proportions of the number of NAPLAN data linkages from Year 3 to Year 9 by jurisdictions

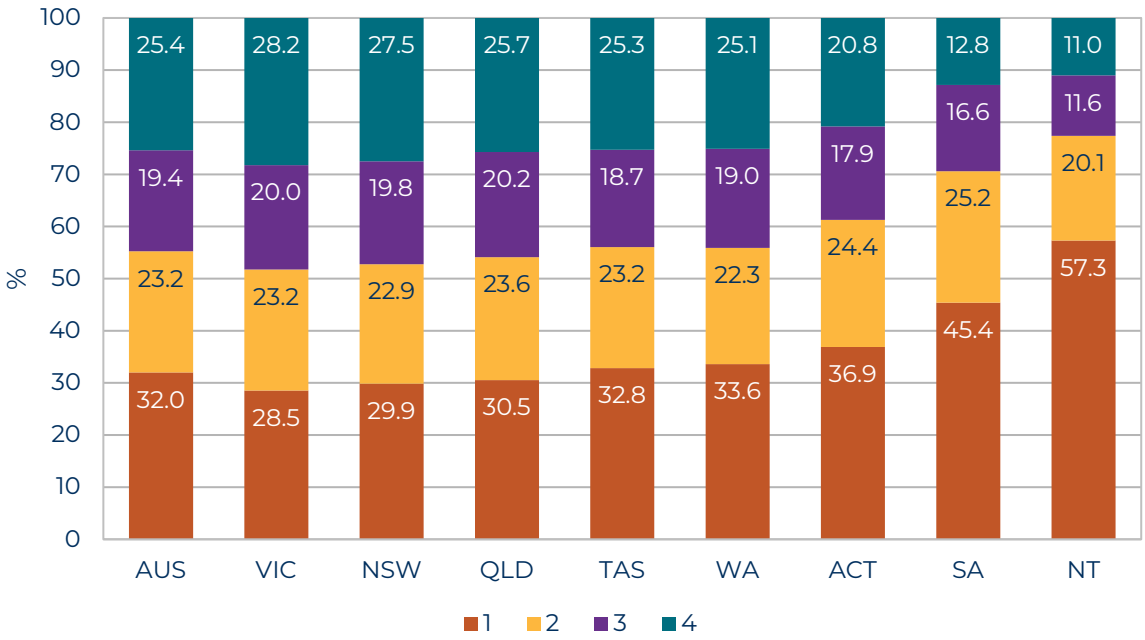


Table 4 contains the rates of matching by Year 3 calendar year. The cells in bold show where the highest proportion of matching was expected to occur. These expectations were based on how many NAPLAN rounds each cohort had experienced. Only 7 Year 3 cohorts (2008 to 2013 and 2015) had progressed to Year 9 at the time of the data linkage, meaning 7 cohorts could be matched completely by tracking the students through from Year 3 to Year 9. The majority of Year 3 students (99.8%) in 2014 could only be tracked to Year 7 due to the cancellation of NAPLAN in 2020. Similarly, the Year 3 students in 2016 and beyond had mostly not reached Year 9 by 2021, meaning they could not be tracked completely through to Year 9.

The observed matching rates corresponded to these expectations and ranged from 78.4% to 100%. The achieved average linkage rate of 87.5% (average of the bolded percentages in Table 4) indicates the effectiveness of the matching strategies that were utilised.

Table 4: Proportions of the number of linked records for Year 3 to Year 9 NAPLAN data by Year 3 calendar year

Number of Year 3 to Year 9 linked records for NAPLAN data (%)				
Year 3 calendar year	**1**	**2**	**3**	**4**
2008	10.9	5.9	4.8	**78.4**
2009	8.4	3.7	4.4	**83.4**
2010	6.5	4.1	4.6	**84.8**
2011	5.1	4.5	4.4	**86.0**
2012	5.0	4.4	6.7	**83.9**
2013	5.0	4.4	9.7	**80.9**

Number of Year 3 to Year 9 linked records for NAPLAN data (%)				
Year 3 calendar year	**1**	**2**	**3**	**4**
2014	5.0	6.1	**88.6**	0.3
2015	5.0	7.9	5.3	**81.8**
2016	6.9	**92.9**	0.1	0.0
2017	8.8	5.9	**85.4**	0.0
2018	**99.8**	0.2	0.0	0.0
2019	8.3	**91.7**	0.0	0.0
2021	**100.0**	0.0	0.0	0.0

Note: Expected matching rates are in bold. Refer to the text for the descriptions of expected matching rates. Year 3 in 2020 did not complete any NAPLAN testing due to the COVID-19 pandemic.

3.4. Linked subset comprising reliable data for fully linkable Year 3 cohorts from 2009 to 2013 and 2015

The Year 3 cohorts in 2009 to 2013 and 2015 had their NAPLAN data available until Year 9 for full matching (N = 1,654,716). In total, 1,380,434 students were able to be fully matched from Year 3 to Year 9 with 4 linked records, resulting in an overall matching rate of 83.4%. The matching rates varied slightly across calendar years. As illustrated in Figure 3, the Year 3 cohorts in 2009 to 2013 and 2015 were fully matched to Year 9 data at a rate of between 81% to 86%. The matching rate for the Year 3 cohort in 2008 was lower due to data quality issues in the first year of NAPLAN, leading to this cohort being excluded from the subset data with full linkage.

Figure 3: Year 3 cohorts' full matching rates until Year 9 by calendar year

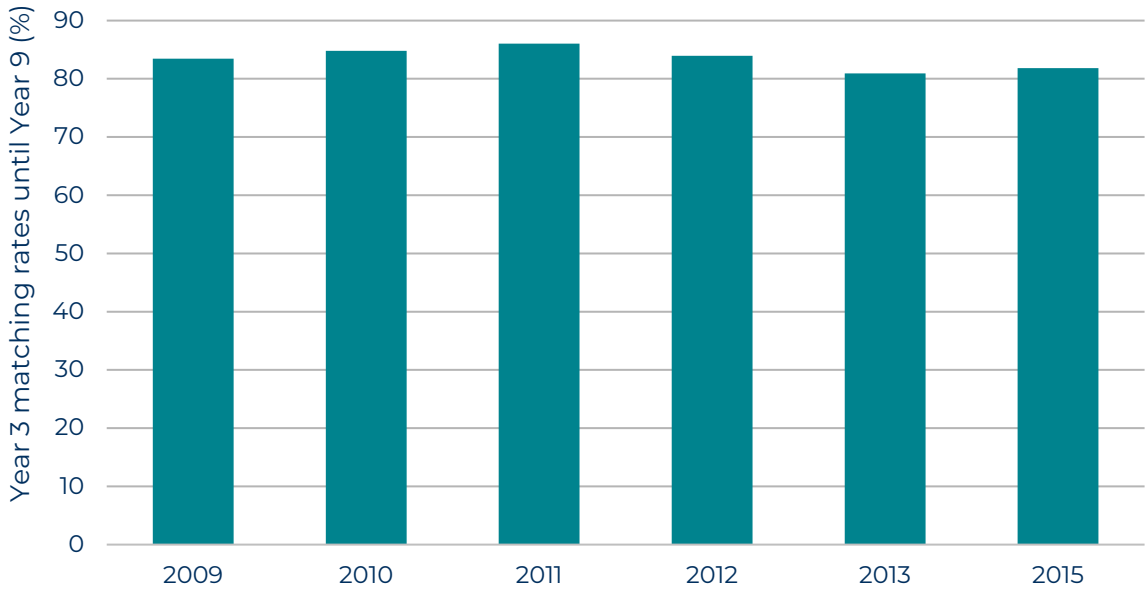


Figure 4 presents the full matching rates by jurisdictions for Year 3 cohorts in 2009 to 2013 and 2015 to their later NAPLAN results up until Year 9. Northern Territory and South Australia showed lower full matching rates mainly because of student ID consistency issues. However, the relatively high number of fully matched students in each state and territory offers statistical opportunities such as creating a stratified national sample (with jurisdiction as a stratum) from the fully matched dataset for use in research. Nonetheless, caution is still needed when any analysis results are interpreted for national patterns and state-level implications.

Figure 4: Full Year 3 to Year 9 matching rates by jurisdiction for Year 3 cohorts in 2009 to 2013 and 2015

We also examined the rates of linkage for the cohorts in the dataset that were not eligible to sit 4 rounds of NAPLAN testing due to their enrolment years. The cohorts of students eligible to have 3 linked records included:

- Year 5 2008 to 2015 and 2017 cohorts who enrolled in Australian schools after mid-Year 3
- the Year 5 cohorts in 2008 and 2009 with missing Year 3 results
- the Year 3 2017 cohort who could not be tracked beyond 2021 to Year 9.

The overall rate of full matching for the number of possible testing rounds for these cohorts was about 81.5%.

Similarly, we explored the rates of linkage for the cohorts who were eligible to have 2 linked records:

- students in Year 7 2008 to 2017 and 2019 cohorts who enrolled in Australian schools after mid-Year 5
- the Year 5 2016 cohort.

The overall matching rate of those 2 possible records for these cohorts was about 82%.

4. Discussion

4.1. Opportunities and potential

The newly linked LLANIA dataset[8] is an important asset for education research, policy, and practice in Australia. More than a decade on from the beginning of NAPLAN, this linkage has realised a significant missing piece of the potential of that project – the ability to track students' performance and growth throughout their schooling.

A natural first application of the LLANIA dataset is the development of empirical trajectories of learning growth. Latent growth modelling, which would require fully matched data from Year 3 to Year 9 for every record, is now possible using a subset of the national data. The potential insights to be gained from this sort of work initially centre on gaining a deeper understanding of how typical students grow in their learning. While these themes are already under exploration in Australia (Larsen et al., 2022), the addition of the LLANIA has the potential to accelerate the establishment of rigorous and representative findings in this space. Following on from this, many extensions of current knowledge are possible through using established average trajectories of learning growth as an evidence base for evaluating the effectiveness of interventions or innovations in teaching and learning compared to current practice.

Further insights are possible for groups of students facing disrupted learning, such as the cohorts moving through school over the course of recent natural disasters in Australia or through the COVID-19 pandemic. Establishing the LLANIA dataset means any changes to the trajectories of learning growth for affected cohorts can be identified accurately through differences from average or non-affected cohorts' trajectories. Research of this kind using the NAEP in the United States is currently coming into publication (Lewis et al., 2022). The newly linked LLANIA dataset is a crucial data resource for understanding the impact of wide-scale events on Australian students.

Similarly, clearer focus can be brought to equity groups in examining how their trajectories of learning growth differ from expected trajectories and the key predictors of this, as has recently been investigated in England using the NPD data (Gorard & Siddiqui, 2019). Exploring this avenue of research in Australia using complete student population data from the LLANIA dataset could deliver much needed evidence supporting clearer direction on policy and practice aimed at closing those gaps.

Insights pertaining to specific levels – such as the student level, the school level, different schooling sectors or geographical regions – can also be improved using the LLANIA dataset. Specifically, multilevel mixed effects models can accommodate all records (regardless of the number of linkages) in estimating learning growth trajectories (Magezi, 2015). Thus, it is possible to maximise the use of the full LLANIA dataset, incorporating students who change schools, as well as new arrivals and departures, across states over the period.

8 We use 'LLANIA dataset' from hereon to refer to both the full dataset with data suppression and masking strategies applied, or the fully de-identified LLANIA national sample.

4.2. Limitations and future directions

Several limitations applied to the data linkage project. The source data did not include tracking of records on student mobility across states and territories. This lack of interstate tracking in addition to the lack of national student identifiers limited the matching of data for students who moved across jurisdictions during their NAPLAN years. Understanding the academic achievement and transitions of this particular group of students might require a different approach, such as the development of new national student identifiers to be used in future national assessments.

The varying matching rates across jurisdictions and calendar years could also present a limitation. However, when we viewed the comparisons of Year 3 students who were fully matched to the total Year 3 population (see Appendix A), there was not a large degree of difference in the demographics or in students' initial Year 3 performance. Nevertheless, it would be desirable to move towards complete matching of student records through future initiatives such as the introduction of a national student identifier.

4.3. Conclusion

The LLANIA dataset – containing the newly linked national longitudinal NAPLAN data of over 6 million students as they moved from Year 3 to Year 9 over the last 13 years – together with the LLANIA national sample is anticipated to support a wide range of future projects, contributing to new knowledge both within theoretical and applied contexts. The expected expansion of evidence and knowledge will significantly benefit Australian students and schools, as policy and practice are informed with greater detail and precision than has been possible to date.

References

Australian Curriculum, Assessment and Reporting Authority. (2022). *ACARA's privacy policy*. Retrieved May 4, 2023, from https://www.acara.edu.au/contact-us/privacy

Australian Institute of Family Studies. (n.d.). *Growing up in Australia: The longitudinal study of Australian children*. Retrieved May 4, 2023, from https://growingupinaustralia.gov.au/

Gorard, S., & Siddiqui, N. (2019). How trajectories of disadvantage help explain school attainment. *Sage Open*, *9*(1). https://doi.org/10.1177/2158244018825171

Goss, P., Sonneman, J., Chisholm, C., & Nelson, L. (2016). *Widening gaps: What NAPLAN tells us about student progress*. Grattan Institute. https://apo.org.au/node/62241

Institute for Fiscal Studies. (n.d.). *National Pupil Database*. Retrieved May 4, 2023, from https://ifs.org.uk/national-pupil-database

Larsen, S. A., Little, C. W., & Coventry, W. L. (2022). The codevelopment of reading and attention from middle childhood to early adolescence: A multivariate latent growth curve study. *Developmental Psychology*, *58*(6), 1017–1034. https://doi.org/10.1037/dev0001344

Lewis, K., Kuhfield, M., Langi, M., Peters, S., & Fahle, E. (2022). *The widening achievement divide during COVID-19*. Center for School and Student Progress. https://www.nwea.org/research/publication/the-widening-achievement-divide-during-covid-19/

Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: An introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00002

National Center for Education Statistics. (2022.) *Intended meaning of NAEP*. Retrieved May 4, 2023, from https://nces.ed.gov/nationsreportcard/guides/

National Centre for Vocational Education Research. (n.d.). *Longitudinal surveys of Australian youth (LSAY)*. Retrieved May 4, 2023, from https://www.lsay.edu.au/

Appendix A: Comparisons across the Year 3 matched cohorts and overall cohorts in 2009 to 2013 and 2015[9]

Figure A1: Comparisons of Parent 1 occupation across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015

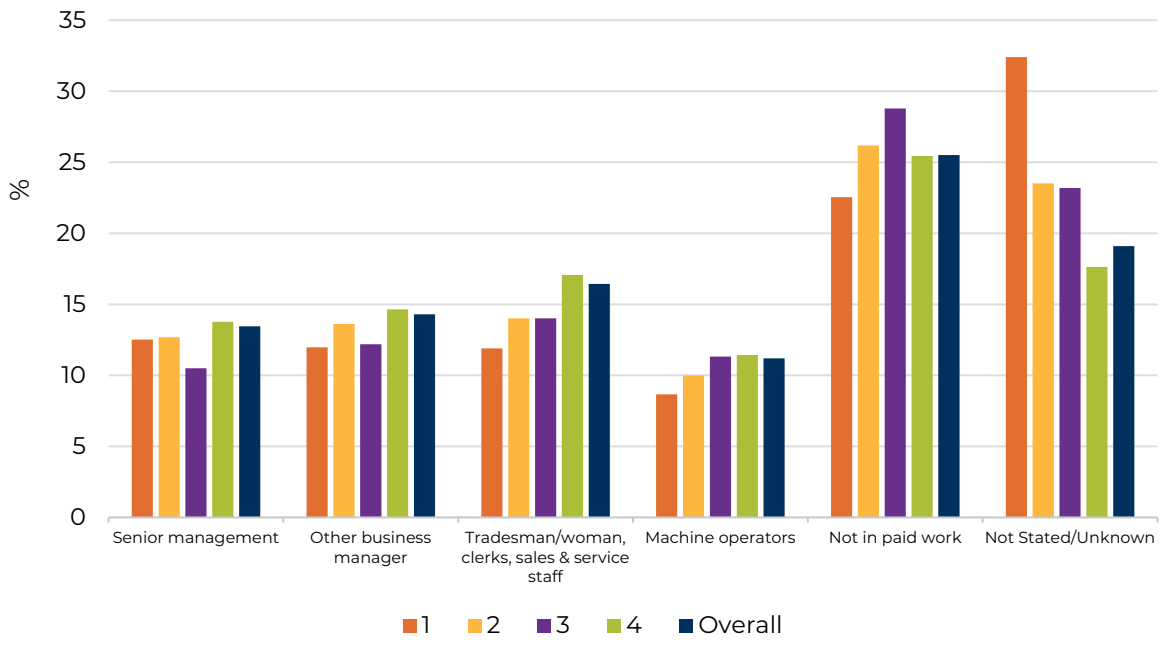
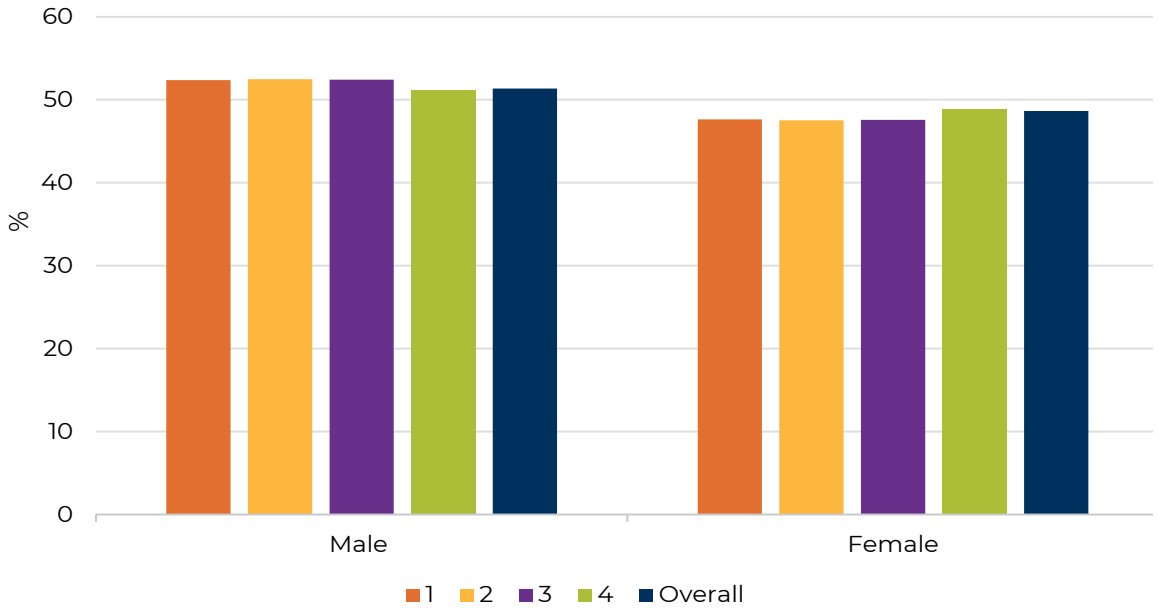


Figure A2: Comparisons of gender across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015



9 All breakdowns presented in this appendix were generated from the full LLANIA dataset prior to the application of data suppression and masking strategies.

Figure A3: Comparisons of First Nations student proportions across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015

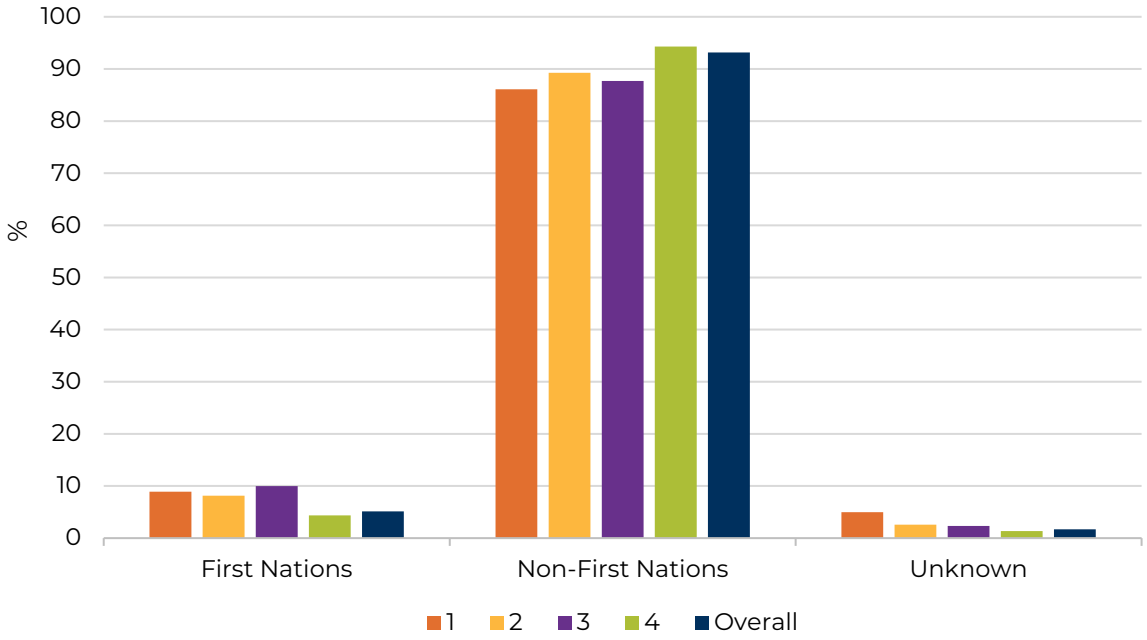


Figure A4: Comparisons of LBOTE student proportions across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015

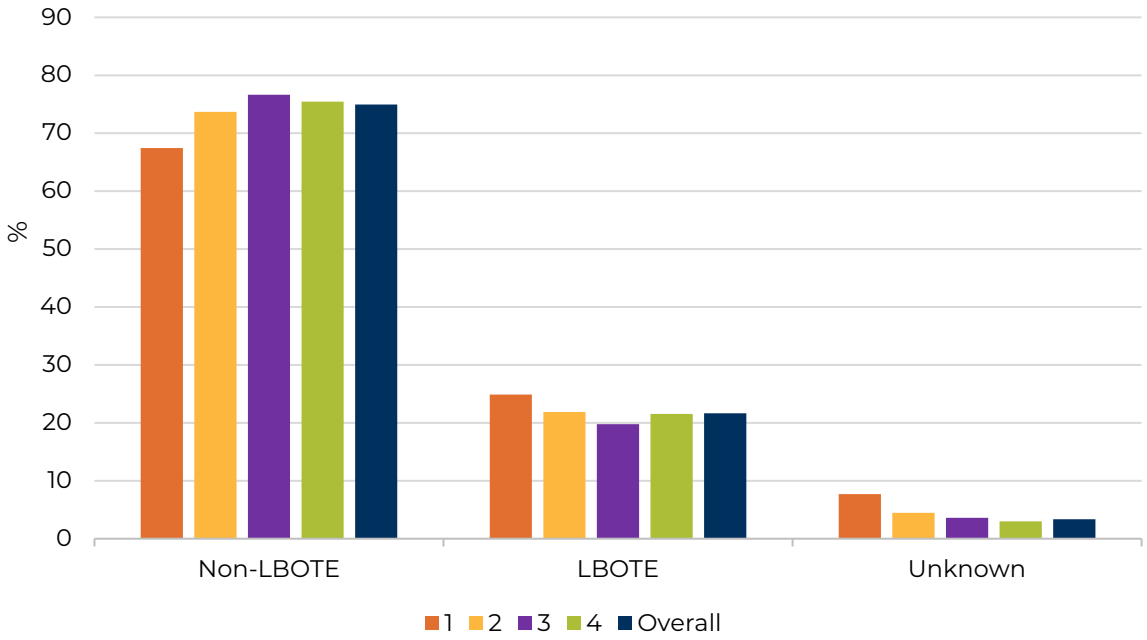


Figure A5: Comparisons of remoteness across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015

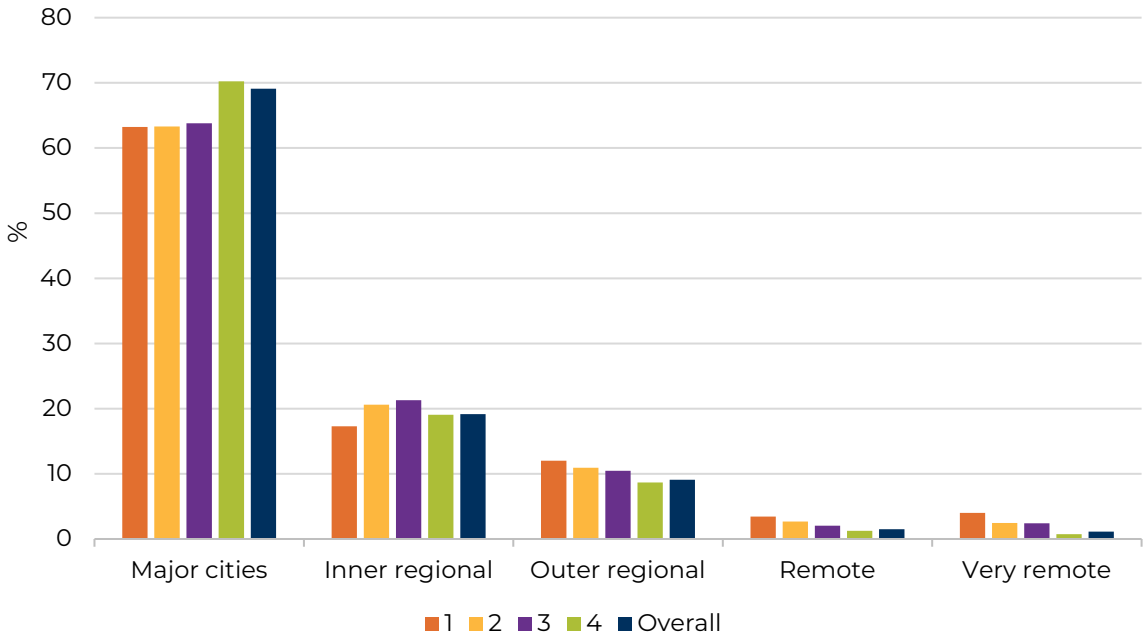


Figure A6: Comparisons of school sector across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015

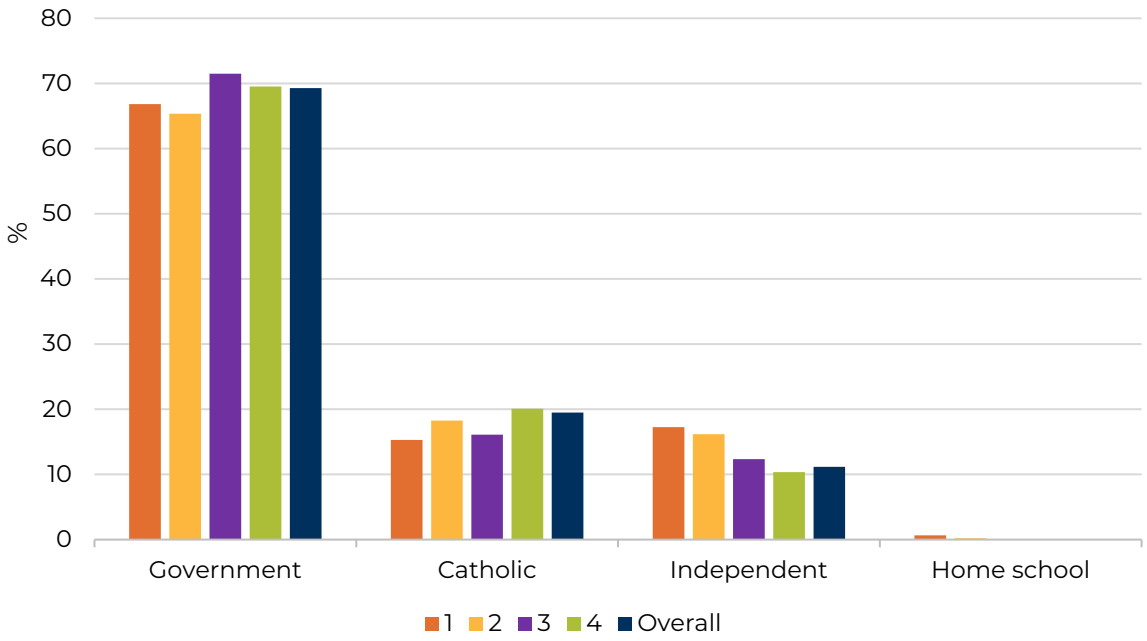
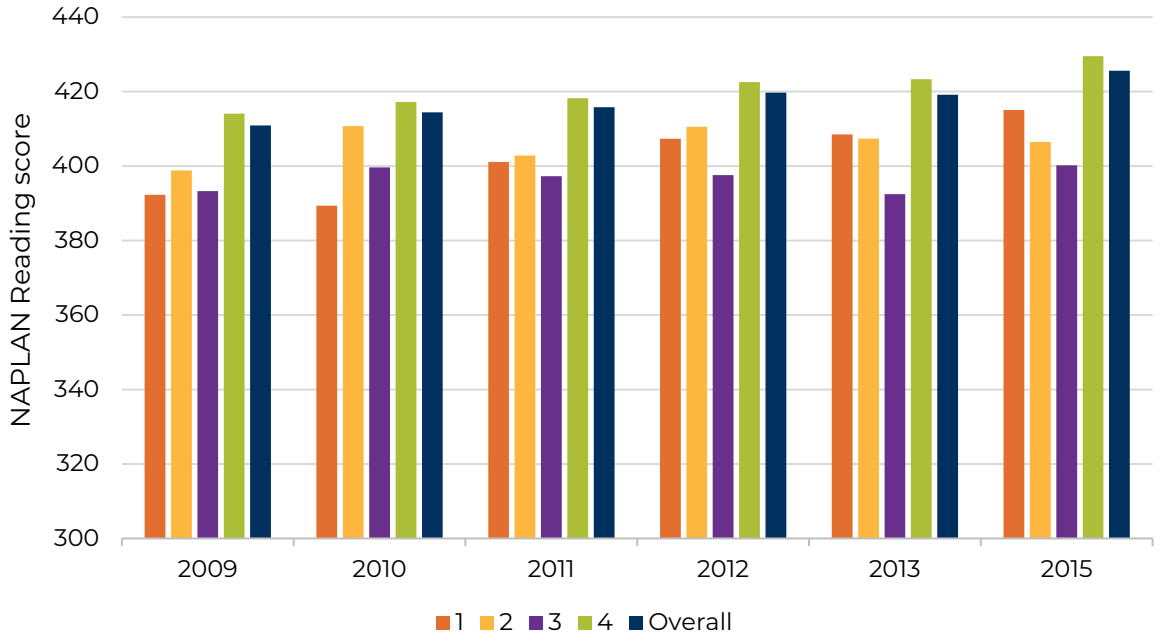


Figure A7: Comparisons of Year 3 reading performance across the Year 3 cohorts with 1 to 4 linked records and the overall cohorts in 2009 to 2013 and 2015



For more information visit
edresearch.edu.au

290224.CH.AA.V2