Received: 24 August 2023 | Revised: 3 May 2024 | Accepted: 7 May 2024

DOI: 10.1002/ajs4.341

ORIGINAL ARTICLE

WILEY

Are Australian students' academic skills declining? Interrogating 25 years of national and international standardised assessment data

Sally A. Larsen

School of Education, University of New England, Armidale, New South Wales, Australia

Correspondence
Sally A. Larsen, School of Education, University of New England, Armidale, NSW 2351, Australia.
Email: slarsen3@une.edu.au

Abstract

Standardised tests of academic basic skills are an established feature of contemporary Australian schooling. Assessment results are widely reported and directly influence educational policymaking. Furthermore, Australian national educational priorities are linked to educational system accountability via the results of standardised tests. Given the influence and importance of assessment data, this paper aimed to collate publicly available data from four assessment programmes undertaken by Australian students, and document long-term trends in average achievement across all available assessments. Results are reported from three international assessments, the Progress in International Reading Literacy Study, the Trends in International Mathematics and Science Study and the Programme for International Student Assessment (PISA), along with the only Australian assessment, the National Assessment Program: Literacy and Numeracy. Of these four, only PISA demonstrated systematic declines in average scores over time. For the remaining three programmes, results in the primary school years showed initial improvements that were subsequently maintained over remaining iterations of the tests. In secondary school, students' average results neither declined nor increased appreciably over time. The consensus of the four largest assessment programmes undertaken by Australian students since

1995 thus fails to support the prevailing narrative of a broadscale decline in academic skills attainment.

KEYWORDS
longitudinal trends, NAPLAN, PIRLS, PISA, TIMSS

1 | INTRODUCTION

Standardised assessments of academic abilities are now part of the fabric of contemporary K-12 schooling, both in Australia and internationally (Verger, Parcerisa, & Fontdevila, 2019). Over the last 40 years, multiple assessment programmes have been developed and administered to Australian schoolchildren. These assessments generally focus on students' attainment of the basic skills that are considered foundational to academic success. Tests assess students' literacy, including reading comprehension, spelling and writing; numeracy, including mathematical content knowledge and applications; and knowledge and applications of science. Results of these tests are subsequently interpreted as evidence of the educational attainment of the population of students at various points of their schooling careers. Test performance is also used as evidence of the capacity of schools and teachers to impart these basic skills to their students (McGaw et al., 2020; Thompson, 2013), to identify areas of concern (e.g., groups of students who underachieve relative to others), and to inform policies for educational reform (Lingard, 2011; Lingard et al., 2014). It is therefore paramount that data generated by these assessments are accurately and holistically interpreted.

While Australian students participate in multiple assessments, results of each are generally interpreted independently of others. To date, a comparison of results on all testing programmes across the span of time that Australian students have participated has not been documented in detail. Given the influence of standardised test results for evaluating school system performance and for informing policy, it is vital data from all the assessments undertaken by Australian students are accessible and reported accurately. The purpose of this paper, therefore, was to compile publicly available data on Australian students' average achievement in the four largest and longest-running national and international standardised assessment programmes. The paper documents trends in average results for each year Australian students participated in each assessment, providing a comparative overview of achievement on these assessments from 1995 to 2022.

1.1 | Standardised assessment programmes in Australia

Data from standardised assessments are increasingly influential in Australian educational policymaking, school improvement programmes and teaching practice (McGaw et al., 2020). The rise of standardised assessments in Australia, and elsewhere, and their influence on educational policy has been extensively discussed (Ball, 2015; Gillis et al., 2016; Lingard, 2011; Lingard & Sellar, 2013; Lingard et al., 2014; Savage et al., 2013). Results of standardised tests are embedded within what Savage et al. describe as "market-based models of governance", which emphasise principals of transparency and accountability (*c.f.* Verger et al., 2019, p. 162). From this perspective, numerical scores on standardised tests provide schools, policymakers and the public with (seemingly) precise and irrefutable information about students' attainment and progress and, by extension, the health of the education system as a whole. Using assessment data, educational outcomes can be easily and quickly compared and evaluated within and between nations, and policy recommendations be made (Cumming et al., 2019;

Gillis et al., 2016; McGaw et al., 2020). It is difficult to argue against the rationality of numbers. That said, standardised tests are a limited representation of school achievement (Wu & Hornsby, 2014) given that they assess a narrow range of educational domains. Irrespective, their influence on Australian educational policy is now far-reaching, outweighing any other kind of evidence on students' school progress.

For Australian governments, standardised assessments are the operational means to answer questions about whether students are attaining the requisite skills to effectively participate in the future workforce, and whether they are developing into active and informed citizens (Department of Education, Skills and Employment [DESE], 2019). For example, the *Measurement Framework for Schooling in Australia* (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2020) links the national educational priorities outlined in the *Alice Springs* (*Mparntwe*) *Education Declaration* (DESE, 2019) to system accountability and policy development via the results of standardised tests. In the language of economic rationalism, student achievement is a Key Performance Measure (KPM), providing "evidence of the outcomes of schooling" (p. 8). If students are not attaining the minimum proficiency standards set jointly by testing authorities and ACARA, then arguably, targeted improvements to schooling can be made. In this view, large-scale, generalisable information about student achievement is essential for effective education policymaking.

Given the central importance of results of standardised tests in evaluating Australian students, teachers, schools and education systems, it is notable that there have been few attempts to compile time series information about student achievement in all existing and ongoing assessment programmes. The National Assessment Program: Literacy and Numeracy (*NAPLAN*) *Review* (McGaw et al., 2020) is one exception. This review reported patterns of achievement in national and international assessment programmes up to 2019, finding varied evidence indicating improvement, decline or minimal change, depending on the assessment programme, age or school year of students tested, and state or territory. While the information reported in the 2020 review is useful, comparing achievement on standardised tests was only one of several aims in the report. An updated and expanded examination of standardised assessment data in the Australian context is warranted for several reasons. First, data from the 2021 round of the *Progress in International Reading Literacy Study* (*PIRLS*) became available in mid-2023, allowing reporting on three waves of PIRLS rather than two. Similarly, the 2019 round of the *Trends in International Mathematics and Science Study* (*TIMSS*) were not available for the 2020 NAPLAN review and are reported in this paper. Data from the latest round of testing in the *Programme for International Student Assessment* (*PISA*) were released in 2023 and are also included. Finally, with the shift to an online, adaptive assessment now complete, the time series for *NAPLAN* data was re-set in 2023 (ACARA, 2023). This means that data from the 2008 to 2022 NAPLAN assessments can be reported in their entirety, with results from 2023 onwards representing a new, slightly different assessment. Full time series data for the school years included in the assessment programme (Years 3, 5, 7 and 9) in the new iteration of NAPLAN will not be available until 2029.

The aim of the current paper, therefore, was to compile information generated by the four national and international educational assessments identified in the *Measurement Framework for Schooling in Australia* (ACARA, 2020) since these assessments are intended to be used to evaluate the progress of Australia's schools and students. I update and extend the chapter in the *NAPLAN Review* (McGaw et al., 2020) by visually presenting a more comprehensive breakdown of each of the assessment programmes. My intention is to provide easily accessible information on the patterns of average student achievement on these assessments for all available test years. Comprehensive information about national progress on all major assessments is crucial if governments intend to rely on these data to inform policy development and evaluate schools and teachers.

1.2 | Standardised assessments undertaken by Australian students

In this section, I provide brief information on each of the four assessments reported in this paper. I describe the focus and content of each test, the sampling method, and how scale scores and achievement bands are generated. Additional and extensive information on each testing programme can be found on the Website for each, in official reports of results, and in technical documentation. For each assessment programme, I cite the official report from the most recent round of assessments included in this paper, though additional historical reports are publicly available.

It is worth noting that the term "basic skills" is taken to mean those fundamental, functional academic skills and capabilities that students are expected obtain as they progress through their schooling career (*c.f.* Organisation for Economic Co-operation and Development [OECD], 2015). Basic skills include, for example, comprehending written passages in different genres, performing basic mathematical operations and completing simple, functional writing tasks. Some standardised tests focus on assessing basic skills (e.g., NAPLAN), while others may put more emphasis on higher order application and interpretation skills (e.g., PISA).

The *Progress in International Reading Literacy Study* (*PIRLS*) assesses the reading skills of students in Year 4 (Hillman et al., 2023). The PIRLS uses a two-stage stratified cluster sampling method to select full classes of students nested within randomly selected Australian schools. Design weights are incorporated if necessary so that while the sample of students undertaking the PIRLS tests is small relative to the full population of Year 4 students, the sample is representative of this population. Australian students have participated in three PIRLS rounds: 2011, 2016 and 2021. The number of schools, students and the average age of students undertaking PIRLS in each participation year is in Table 1.

The PIRLS tests assess students' understanding of the "purposes for reading" and the "processes of comprehension" (Hillman et al., 2023, p. 2). Tests comprise one literary text and one information text, and questions are in multiple-choice or short-answer format. Item response theory methods are used to transform students' raw scores to the PIRLS historical scale (developed in 2001). The PIRLS scale has a mean of 500 and a standard deviation of 100 points. Average scores are reported for the whole sample and for subgroups, and student achievement is also reported as meeting one of four proficiency bands: *Low*, *Intermediate*, *High* and *Advanced*. To meet Australia's National Proficiency Standard, students should meet or exceed the *Intermediate* PIRLS benchmark. Additional information about PIRLS assessments can be found on the Australian PIRLS Website, and all data used in this study are publicly available (https://www.acer.org/au/pirls/reports-and-data).

The *Trends in International Mathematics and Science Study* (*TIMSS*) assesses maths and science achievement in Year 4 and Year 8. Students are sampled using the same methodology applied in PIRLS tests (described above), and Australian students have participated in six rounds of TIMSS in total, beginning in 1995, missing the 1999 round and then the five rounds

TABLE 1 Total number of Australian schools and students, and average age of students participating in the PIRLS assessments in Year 4 in each year.

	2011	2016	2021
Year 4			
Schools	280	286	281
Students	6126	6341	5487
Average age	10	10	10

TABLE 2 Total number of Australian schools and students, and average age of students participating in the TIMSS assessments in Year 4 and Year 8 at each year.

	1995	2003	2007	2011	2015	2019
Year 4						
Schools	191	204	229	280	287	287
Students	6507	4675	4108	6146	6057	5890
Average age	9.5[a]	9.9	9.9	10	10	10.1
Year 8						
Schools	158	210	228	275	285	284
Students	6196	5355	4069	7556	10,338	9060
Average age	13.5[a]	13.9	13.9	14	14	14.1

[a]Mean age is approximated from the information on the 1995 samples reported in Gonzales and Foy (1997).

every four years from 2003 to 2019. The number of schools and students participating in each round, and the average age of participating students is in Table 2.

The test documentation for TIMSS indicates that the tests are "designed, broadly, to align with the mathematics and science curricula used in the participating education systems and countries" (Thomson et al., 2020). For mathematics, Year 4 students are assessed on three content domains including number, measurement and geometry; in Year 8, number and geometry content domains are retained, with algebra and data, and probability added to make four domains. For science, Year 4 students are assessed on three content domains: life science, physical science and earth science; Year 8 students are also assessed on earth science, along with biology, chemistry and physics. The TIMSS tests are designed to assess subject matter knowledge (the content dimension) and students' ability to use knowledge, application and reasoning in their response to questions (the cognitive dimension). Test question format is either multiple-choice or short-answer. While TIMSS has begun the process of shifting to an online test format, Australian students in 2019 (and all previous years) completed paper-based tests (Thomson et al., 2020).

For each test, item response theory is used to generate a scale with a mean of 500 and a standard deviation of 100 points (Thomson et al., 2020). Each new test is mapped to the historic TIMSS scale so that mean changes over time can be interpreted. However, each test should be interpreted with reference only to other iterations of the same test. That is, comparisons cannot be made between Year 4 and Year 8 same-domain tests nor between different domain tests at the same year. Student scores on tests are subsequently mapped to one of four proficiency bands: *Low*, *Intermediate*, *High* and *Advanced*. Students who meet or exceed the *Intermediate* benchmark are also meeting Australia's National Proficiency Standard. All official reports and data used in this study are available on the TIMSS Website (https://www.acer.org/au/timss).

The *Programme for International Student Assessment* (*PISA*) is perhaps the best known of the international assessments undertaken by Australian students. Every three years since 2000, representative samples of 15-year-old Australian students have been assessed, with the 2022 PISA round, making the eighth cohort to participate in the three-yearly cycle. Interruptions caused by the COVID-19 pandemic meant the round scheduled for 2021 was pushed back to 2022, with initial results released in late 2023 (Australian Government, 2023a). A stratified random sampling approach is used to select a representative sample of schools, and a random sample of students within those schools then sits the tests (Thomson et al., 2019). As for PIRLS and TIMSS, while the absolute number of students sitting the PISA tests is small compared with the population, the sampling methodology is robust and the samples can be considered representative of the population of 15-year-old students in Australia. Students are sampled based on their age rather than their

TABLE 3 Total number of Australian schools and students, and birth date range of sampled students participating in the PISA assessments in each year.

	2000	2003	2006	2009	2012	2015	2018	2022
Schools	231	321	356	353	775	758	740	743
Students	5176	12,551	14,170	14,251	14,481	14,530	14,273	13,437
Birthdate range	1 May 1984–30 April 1985	1 May 1987–30 April 1988	1 May 1990–30 April 1991	1 May 1993–30 April 1994	1 May 1996–30 April 1997	1 May 1999–30 April 2000	1 May 2002–30 April 2003	1 May 2006–30 April 2007

school year, so students in Year 9, Year 10 or Year 11 are included in the samples (Ainley et al., 2020). The numbers of schools and students and the birthdate range for participant selection in each round of PISA are reported in Table 3.

The PISA tests focus on the application of knowledge and skills in three subdomain tests: Reading Literacy, Mathematical Literacy and Scientific Literacy. Unlike TIMSS, PIRLS and NAPLAN (see below), PISA tests are not specifically mapped to the content of the Australian curriculum (McGaw et al., 2020). Instead, PISA aims to:

> …measure the cumulative outcomes of education by assessing how well 15-year-olds … are prepared to use the knowledge and skills in particular areas to meet real-life opportunities and challenges.
>
> (Thomson et al., 2019, p. xiii)

The application of knowledge and skills is therefore central to the purpose of PISA tests and guides their construction: The term "literacy" for each domain denotes this purpose. That is, unlike the other standardised assessments examined in this paper, PISA does not assess basic skills alone, rather the intention is to assess how well students can apply knowledge from the specific tested domains to solve problems.

The format of the tests is generally a stimulus text followed by two or more multiple-choice or short-answer questions (Thomson et al., 2019). Mathematical and Scientific Literacy tests contain stimulus items focussing on problems or scenarios requiring knowledge or understanding of concepts relevant to those domains. Reading Literacy tests contain a variety of texts for different purposes (e.g., descriptions, narrations, expositions, arguments and instructions) and assess skills including fluency, locating information, understanding and evaluation.

Item response theory is used to generate scales for each test domain separately. Each scale has a mean of 500 and standard deviation of 100, and scores are subsequently mapped to proficiency bands. Each domain is mapped to its historical scale using a common-item equating procedure so that cohort changes can be observed (Thomson et al., 2019). The proficiency bands were developed, and the scales refined, on the first occasion that an assessment domain was the focus area of PISA, that is, Reading Literacy in 2000, Mathematical Literacy in 2003 and Scientific Literacy in 2006. For this reason, proportions of students falling into each proficiency band in each domain are only available from these years forward, while mean scores are available for every year (i.e., 2000–2022). For each domain, the National Proficient Standard for Australian students is set at Level 3 or above. Official reports and all data used in this study are publicly available (https://www.acer.org/au/pisa).

The *NAPLAN* commenced in Australia in 2008. This programme aims to assess the literacy and numeracy attainment and progress of all students from middle primary to middle secondary school. Students sit NAPLAN tests in Reading Comprehension, Writing, Spelling, Grammar and Punctuation, and Numeracy in four biennial school years, Years 3, 5, 7 and 9. This national programme replaced previous state-based standardised assessments, such as the Basic Skills Test (NSW), which began in 1989, and the Western Australian Literacy and Numeracy Assessment, which began in 1998. Unlike the international test programmes, which select representative samples of students to participate in the assessments, NAPLAN is a population census of student achievement, and the only standardised assessment that all students undertake at multiple time points. Approximately one million Australian students sit NAPLAN tests each year with similar proportions across the four assessed year levels (i.e., ~250,000 students in each of Year 3, 5, 7 and 9; ACARA, 2021). The age of students when they sit the tests depends on (a) school entry policies for the state influencing the age at which students entered school, (b) whether students completed a prep or kindergarten year or entered in Year 1 and (c) whether students repeated or skipped a year (though these cases are rare in Australia, approximately 1–2 per cent of students repeat a year; Anderson & Anderson, 2020;

Larsen et al., 2021). Up to 2022, NAPLAN tests were undertaken on the same dates each year, so students were exactly two years older at each NAPLAN test subsequent to Year 3 (average age 8 years 7 months approx., ACARA, 2021).

The tests are scaled using Rasch item response theory, producing a ratio interval-scale (ACARA, 2022). For each assessment, students' results are reported on 1–1000 point scale. Up to the 2022 round of assessments, scale scores were mapped to one of 10 achievement bands. Students' scores in lower school years spanned lower achievement bands and those in higher school years spanned higher bands; for example, Year 3 results spanned Band 1 to Band 6, increasing to a span of Band 5 to Band 10 in Year 9. An equating process allowed each year's assessment to be mapped to the historical achievement scale set in 2008 so that cohort changes could be represented (ACARA, 2021). In addition, achievement across the four years' assessments was horizontally equated so that within-cohort growth could be mapped from Year 3 to Year 9.

The design of NAPLAN means that population achievement can be examined in two ways: (a) between-cohort achievement trends for each year level assessed (i.e., average achievement differences for all cohorts completing Year 3 tests and Year 5 tests), and (b) longitudinal within-cohort trends over the four assessments (i.e., Year 3, 5, 7 and 9). One advantage of NAPLAN tests, compared with the international testing programmes outlined above, is their capacity to track students' achievement over time. This is not possible with cross-sectional assessments occurring at only one school year or age, as is the case for all other assessments reported in this paper.

The NAPLAN content is aligned with the Australian curriculum for English (Reading, Writing, Spelling, Grammar and Punctuation tests) and Mathematics (Numeracy test). For all tests except Writing, students respond to prompts or questions in either a multiple-choice or short-answer format. The Writing tests asks students to respond to a prompt in either a persuasive or narrative genre. Prior to 2011, the genre varied year to year, making comparisons over time difficult. Since 2011, every Writing test, except for 2016, has been a persuasive prompt. Students' scripts are marked against the same criteria each year, and the 2016 narrative task results were later mapped to the persuasive criteria, allowing cross-time comparisons from 2011 to 2022 (McGaw et al., 2020). Additional detail on test items, duration and response formats for instance, can be found in technical reports (ACARA, 2022) and all data used in this study are publicly available (https://www.acara.edu.au/reporting/national-report-on-schooling-in-australia/naplan-national-report). Past test papers for the Years 2012 to 2016 are also available online.

2 | METHODS

The publicly reported mean scores and proportions of students falling into achievement bands or levels for each assessment programme were collated manually in MS Excel, and all figures were generated using the Excel built-in tools. Data from each assessment programme are presented visually such that trends in average achievement over time can be observed for each subdomain of each assessment programme. I acknowledge that plotting average scores for groups of students is a fairly one-dimensional approach to evaluating students' academic achievement, since variability is not considered. Nonetheless, average scores are the primary means of communicating information to the public, and to governments, about students' achievement on all the included tests. Additional figures also show the proportions of students falling in each achievement band for every available round of each assessment. This strategy provides some insights into the spread of achievement across the distribution, though I rely on the cut-off scores set by testing authorities that categorise students into achievement bands (i.e., continuous variables are categorised using rules specified for each assessment), and I do

not use inferential statistics to test whether changes in proportions of students in each band are significantly different from zero. Notwithstanding these limitations, the intention here is to provide broad information about population trends and within-test plots can achieve this aim.

While statistical significance of between-group achievement differences is often documented in national reporting of standardised assessments (McGaw et al., 2020), interpreting statistical significance in this way can be problematic. All of these programmes assess large groups of students (i.e., not fewer than 4000 and up to 260,000 in any assessment), so differences between-cohort groups taking the tests in different years can be statistically significant without being practically meaningful. Furthermore, one aim of interpreting statistical significance is to generalise to a theoretical population. When assessments are designed to assess the entire population of students, as with NAPLAN, statistical significance is arguably meaningless (though comparisons of cohorts may be defensible) (Cowger, 1984).

Alternatively, effect sizes may be selected as the means for interpreting cohort achievement differences within each assessment programme. While effect sizes are less affected by sample size, for cross-sectional assessments, they can only be interpreted for the same test in the same age group. In this paper, I do not make direct comparisons of changes in mean scores or effect sizes between different assessments for two reasons. First, the amount of progress made on standardised assessments over a given time span is greater for younger students than for older (Hill et al., 2008; Kraft, 2020); thus, an effect size difference of 0.2*SD* for students aged 15 has a different interpretation to the same effect size for students aged 7. Comparisons of effect size differences for students of different ages are extremely problematic (e.g., see Australian Education Research Organisation [AERO], 2023 for an example of poor comparisons of effect size differences). Second, even if assessments are vertically equated (as for NAPLAN), the types of abilities tested in similarly named assessments may be qualitatively different in different age groups (Paris, 2005). That is, just because a test is a 'reading' test, the nature of the test will be different for children in Year 3 compared with adolescents in Year 9. This restriction is echoed in Wiliam's (2019) exhortation to meta-analysis of education research to ensure the reported effect sizes are comparing the same things.

For NAPLAN, I focus mainly on cohort differences within-school years (e.g., comparing Year 3 achievement in 2012 with Year 3 achievement in 2017). However, I do also examine proportions of students falling in achievement bands over the four assessed years. For this latter analysis, I rely on the test-equating procedure outlined in the NAPLAN Technical papers (ACARA, 2022), which allows a comparison of scale scores (and hence bands) across year-levels. Nonetheless, I acknowledge that the underlying assumption of vertically scaled tests (such as NAPLAN), that the tests measure the *same theoretical domain* at each assessment point (Briggs & Weeks, 2009), is a fairly strong assumption to make for an assessment programme spanning students aged 8–14.

3 | RESULTS

3.1 | PIRLS

Figure 1 shows mean scores for the Australian sample in the Year 4 PIRLS Reading assessments for 2011, 2016 and 2021. To allow an understanding of the possible variability in PIRLS Reading scores, the *y*-axis is centred at the mean (500), allowing a span of one standard deviation on the historic PIRLS scale (i.e., 100 scale scores = 1*SD*). The average scale score for Australian students increased from 527 in 2011 to 544 in 2016, and 540 in 2021. The improvement from 2011 to the latter two years is approximately +0.15*SD*. The difference between 2016 and 2021 is negligible (Hillman et al., 2023). Figure 2 shows the percentage of Australian students meeting each of the five international PIRLS benchmarks, from "Below low" to "Advanced."

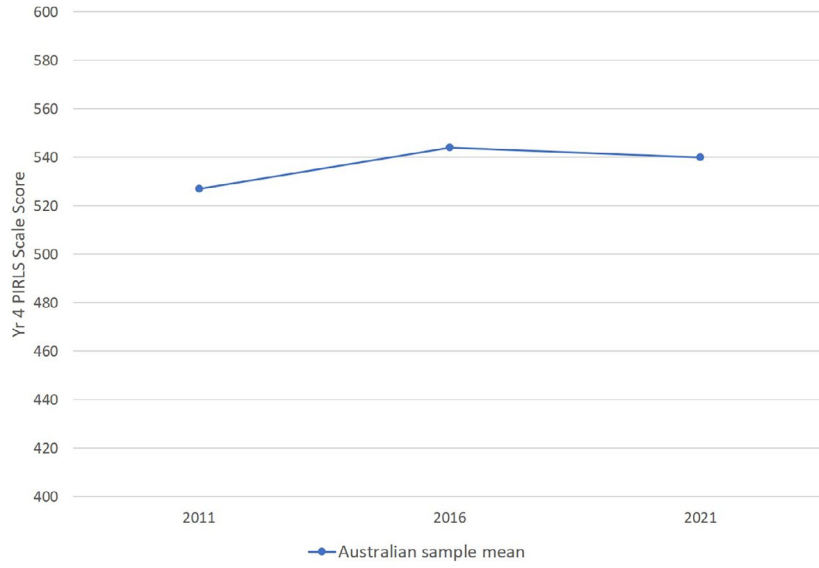

FIGURE 1 Time series showing mean scores in the Year 4 PIRLS Reading assessments for the representative sample of participating Australian students 2011–2021.

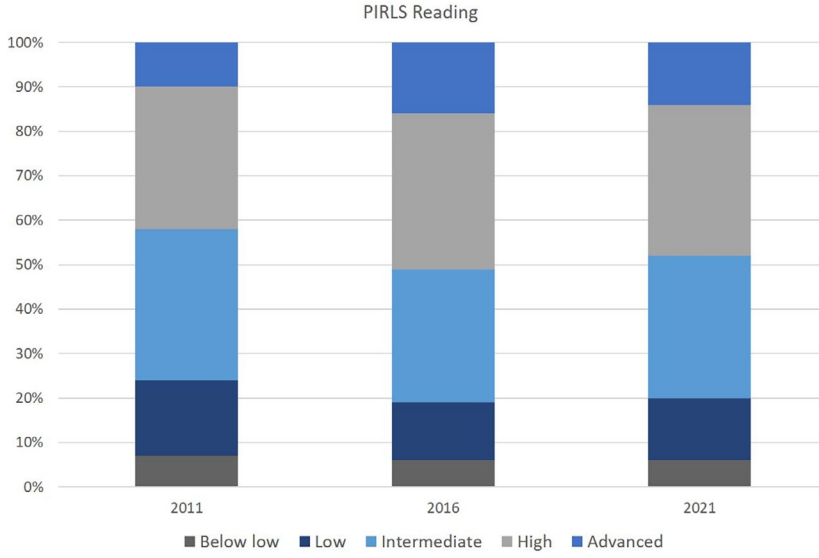

FIGURE 2 Proportions of Australian students meeting each of five benchmarks in Year 4 PIRLS Reading assessments. N.B. The Intermediate benchmark is the Australian minimum standard.

The percentage of students meeting the Advanced and High benchmarks increased from 42 per cent in 2011 to 52/48 per cent in 2016 and 2021, respectively. Conversely, the percentage of students not meeting the Australian minimum standard (Low or Below low benchmarks) reduced from 24 per cent in 2011 to 19/20 per cent in 2016 and 2021. These results suggest a positive trend overall in that reading achievement of Australian Year 4 students improved from 2011 to 2016 and remained similar to 2021, as measured by both the mean performance and the proportions of students attaining the minimum benchmark.

3.2 | TIMSS

Figure 3 shows mean scores in the Year 4 TIMSS Mathematics (top panel) and Science tests (bottom panel) for the six years Australian students have participated in these assessments (1995–2019). For both the Year 4 and Year 8 (Figure 5) TIMSS assessments, the *y*-axis is centred at the historic mean of 500 on the TIMSS scales, with one standard deviation above and below the mean indicating the range of variation in scores ($1SD = 100$). In the Year 4 Mathematics tests, the average achievement of Australian students increased between 1995 and 2007, by 21 scale scores (approximately $0.2SD$). Thereafter, the Australian average Year 4 TIMSS maths achievement has remained relatively stable. For the TIMSS Science assessments (Figure 2,

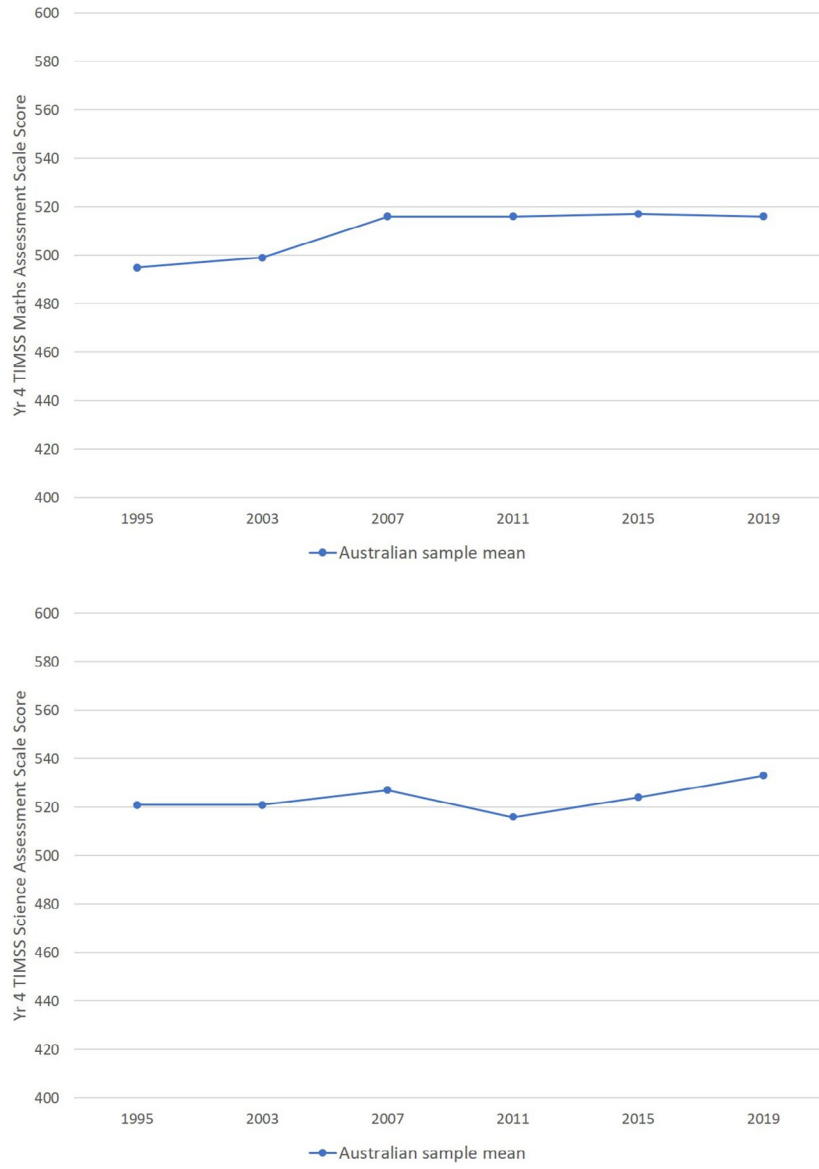

FIGURE 3 Time series showing mean scores in the Year 4 TIMSS Maths assessments (top panel) and Year 4 TIMSS Science assessments (bottom panel) for the representative sample of participating Australian students, 1995–2019.

bottom panel), Australian students performed similarly from 1995 to 2015 with averages ~20 scale scores above the historic mean of 500 (approximately 0.2*SD*). The highest average was from the most recent round of assessments in 2019 (533 on the TIMSS science scale).

Figure 4 shows the proportions of Year 4 students at each assessment year meeting each of the five international benchmarks set by the TIMSS test developers. The Intermediate benchmark is the national minimum standard in Australia. Students falling into the Low and Below low benchmarks are considered to not be meeting expectations for Maths or Science in Year 4. For the Maths tests, between 1995 and 2007, the proportion of students working at the Advanced level increased and thereafter remained stable to 2019. Conversely, the proportion of students not meeting the minimum standard (those in the Low and Below low levels) reduced from almost 40 per cent in 1995 to ~30 per cent from 2011 onwards. There is no strong pattern

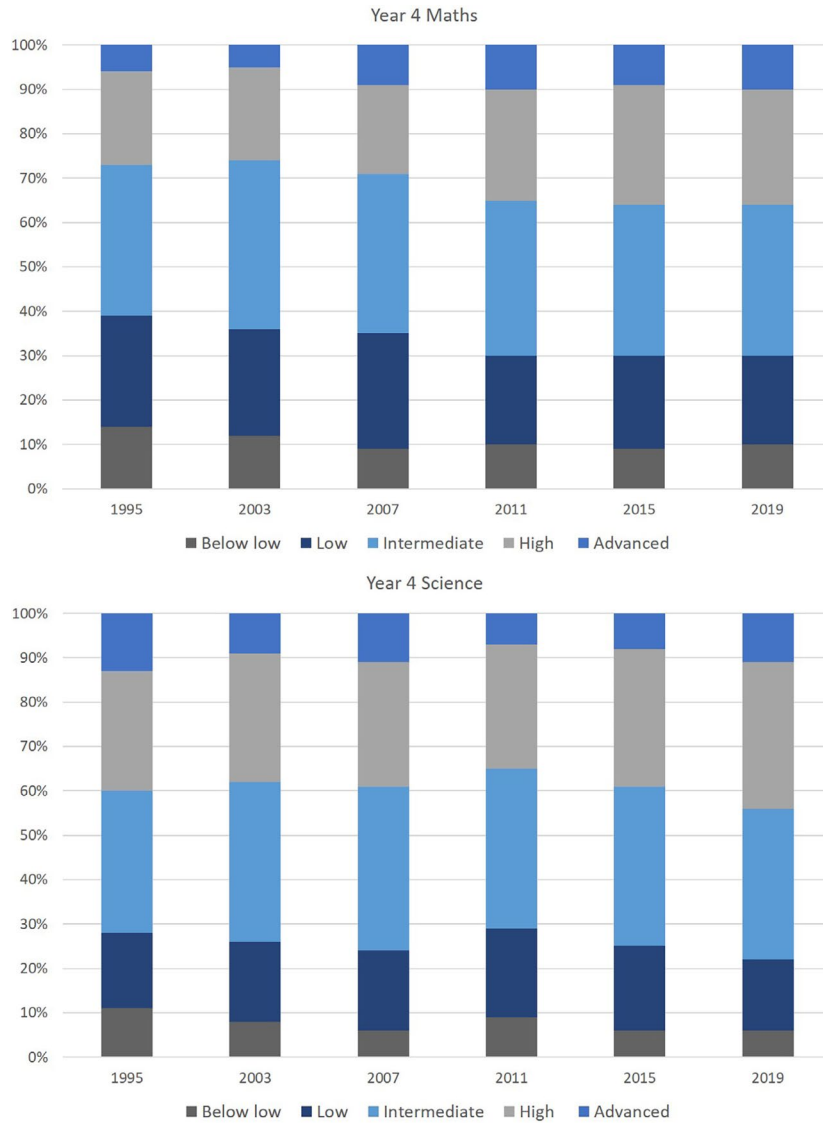

FIGURE 4 Proportions of Australian students meeting each of five achievement benchmarks in Year 4 TIMSS Maths (top panel) and Science (bottom panel). N.B. The Intermediate benchmark is the Australian minimum standard.

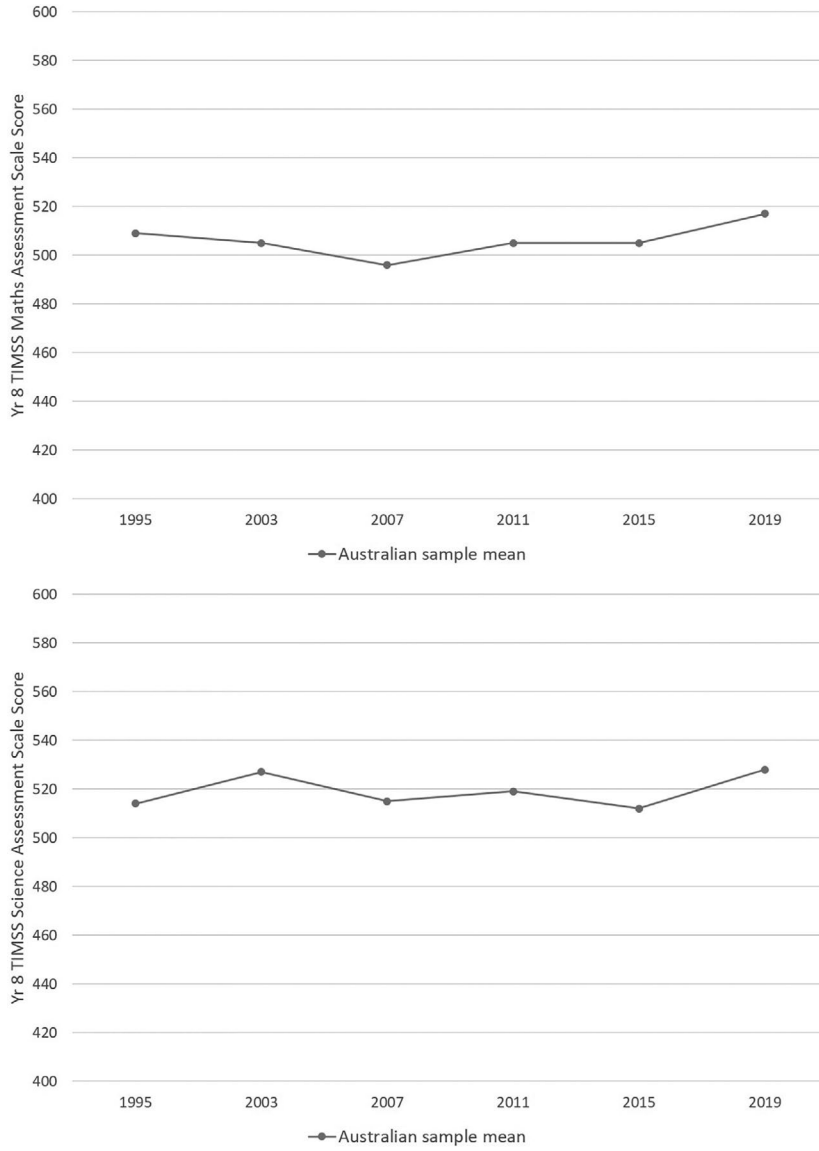

FIGURE 5 Time series showing mean scores in the Year 8 TIMSS Maths assessments (top panel) and Year 8 TIMSS Science assessments (bottom panel) for the representative sample of participating Australian students, 1995–2019.

for the proportions of students at the top and bottom of the achievement distributions for Year 4 Science. Less than 30 per cent of students are categorised as Low or Below low proficiency in Science at all assessments, and the proportion of students meeting the Advanced level fluctuates around 10 per cent.

Average scores for the Australian samples of students undertaking Year 8 TIMSS Mathematics (top panel) and Science assessments (bottom panel) are shown in Figure 5. As with the Year 4 TIMSS tests, Australian Year 8 students did not participate in TIMSS tests in 1998. For Year 8 Mathematics, Figure 5 shows Australian students' average achievement ranging around the midpoint of the scale: 500 scale scores. Students had the highest average achievement in the most recent assessment in 2019, 517 on the TIMSS scale. This represents

approximately 0.2*SD* improvement from the lowest average in 2007 (496 scale scores). Average achievement in Year 8 TIMSS Science tests shows a similar pattern to that of the Year 4 Science tests with the Australian average remaining approximately 0.2*SD* above the scale midpoint at each test year. The highest mean for Year 8 Science was also in the most recent round of assessments, 2019, a score of 528 on the TIMSS science scale (0.25*SD* above the historic mean).

Figure 6 shows the proportions of students meeting each of the five proficiency benchmarks for TIMSS Year 8 Maths (top panel) and Science tests (bottom panel) over all six participation years. The percentage of students meeting the Advanced benchmark has slightly increased from 1995 (7 per cent of students) to 2019 (11 per cent of students), while those in the Below low category have remained stable at around 10 per cent. The proportion of students meeting the Australian minimum standard (Intermediate) declined and then improved between 1995 and

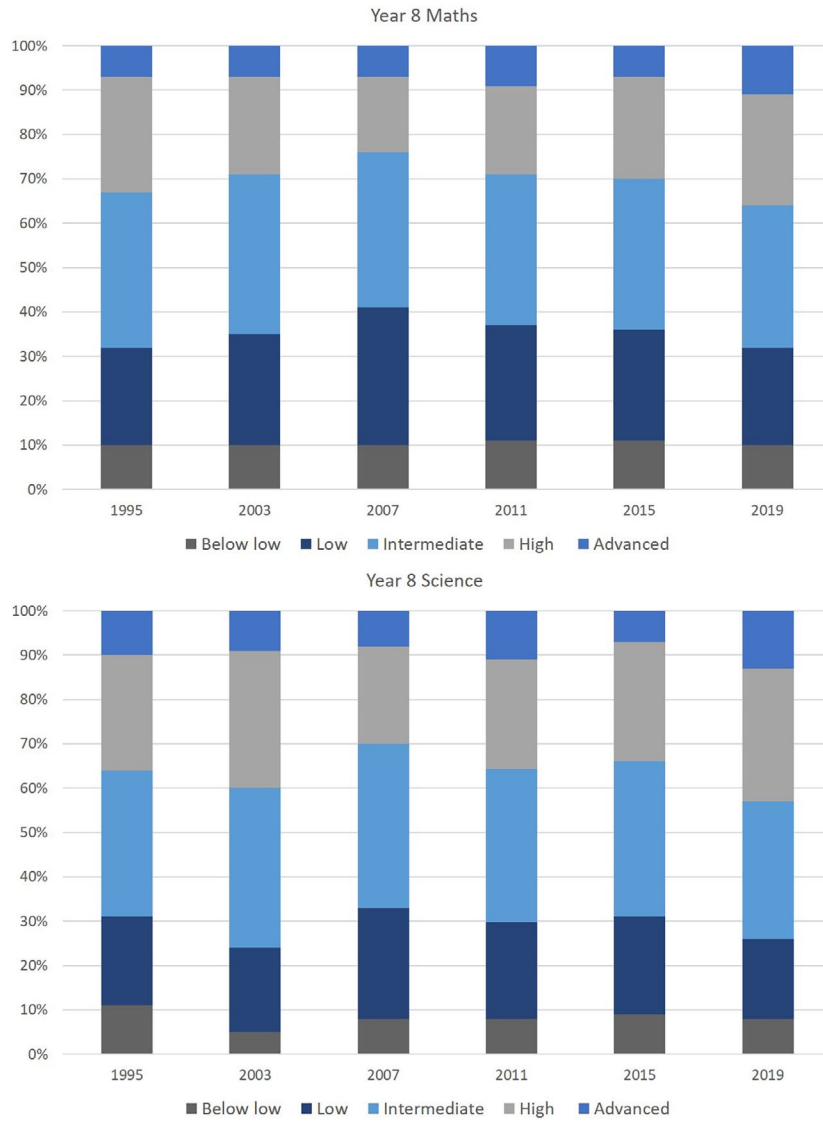

FIGURE 6 Proportions of Australian students meeting each of five achievement benchmarks in Year 8 TIMSS Maths (top panel) and Science (bottom panel). N.B. The Intermediate benchmark is the Australian minimum standard.

2019 such that the same proportion of students did not meet the minimum in both the 1995 and 2019 round of assessments (32 per cent). It is worth noting that the year with the lowest average score (2007) was also the year in which the highest proportion of students did not meet the minimum standard (41 per cent).

Proportions of students falling in each proficiency band fluctuated from year to year for the Year 8 Science tests. Around 30 per cent of students did not meet the minimum standard in each assessed year, though the 2019 round showed some improvement with this proportion declining to 26 per cent. Similarly, those attaining the Advanced benchmark ranged around 10 per cent, with the highest proportion in this band also in the 2019 assessment round (13 per cent). The 2019 overall mean was also the highest of all the assessment years.

3.3 | PISA

In this section, I present data from PISA Reading, Mathematics and Science Literacy tests for all years that Australian students have participated (2000–2022). Mean scores for each round are plotted separately for each domain. Benchmarks for each of the assessed domains were developed in the 2000 round (for Reading), the 2003 round (for Mathematics) and the 2006 round (for Science). The OECD used these years because the respective domain was the focus of the assessment programme in the related year, and students answered a more comprehensive battery of questions for the given domain. These detailed assessments were then used as benchmarks, with future assessments mapped to the benchmarks in the scale equating process in each subsequent year.

There are six proficiency levels for Reading Literacy and seven proficiency levels for Mathematics and Science. The levels for the latter two domains were initially developed with seven categories ranging from "Below Level 1" to "Level 6." For the Reading Literacy domain, six levels were initially set with "Level 5" as the highest. From 2009, PISA included two additional categories at each extreme of the distribution for Reading Literacy, including "Level 6" at the top end, and splitting "Below Level 1" into two categories. However, the cut-off scores on the PISA scale, against which all years' assessments are mapped, were not altered, allowing us to maintain consistency and report the original six levels at each year. In Figure 8, "Level 5" includes the small proportion of students achieving at the later-added Level 6, and "Below Level 1" comprises both of the lowest categories (renamed "Level 1a" and "Level 1b" in 2009).

Figure 7 shows mean scores in the PISA Reading Literacy assessments for the representative Australian sample of students who participated in the tests from 2000 to 2022. The *y*-axis is scaled such that the historical PISA scale mean (500 points) is the centre point, and a span of one standard deviation above and below the mean is represented (i.e., $1SD = 100$ points). The average scores of Australian students have steadily declined since 2000 from an initial score of 528 to an average of 498 in 2022. This decline can be understood as approximately $0.3SD$ on the PISA Reading Literacy scale.

Proportions of Australian students meeting each of six proficiency levels in Reading Literacy is shown in Figure 8. The most notable difference between the initial two assessment years and the subsequent assessment years is the steady increase in the proportion of students categorised as Low or Below Low. In 2000 and 2003, 12 per cent and 9 per cent of students (respectively) achieved in these bottom two levels; this percentage increase to 21 per cent in 2022. Essentially, this is a doubling of the proportion of students "unable to demonstrate the capacity to use their reading literacy skills to solve a wide range of practical problems" (De Bortoli et al., 2023, p. 163). The proportion of students in the highest proficiency level declined from a high of 18 per cent in 2000 to fluctuate around 10–13 per cent of students attaining Level 5 and above in all subsequent years.

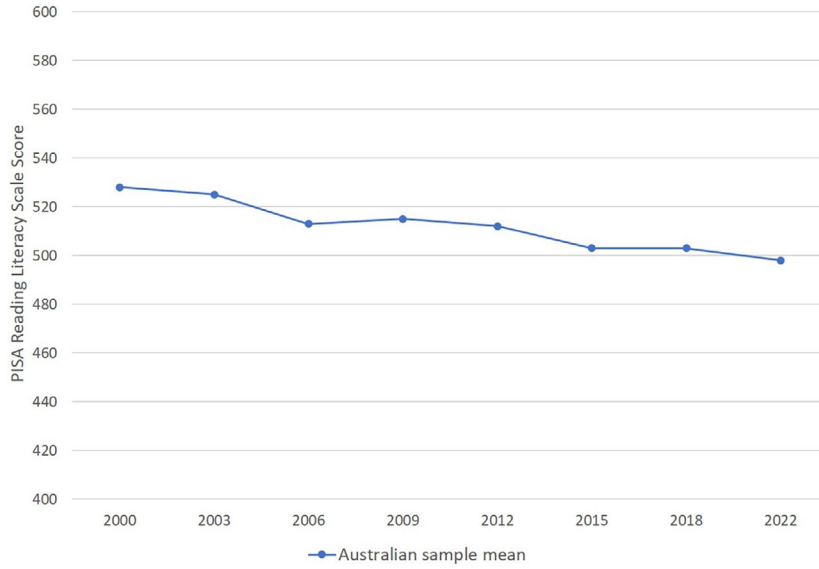

FIGURE 7 Time series showing PISA Reading Literacy test mean scores for the representative sample of participating Australian students 2000–2022.

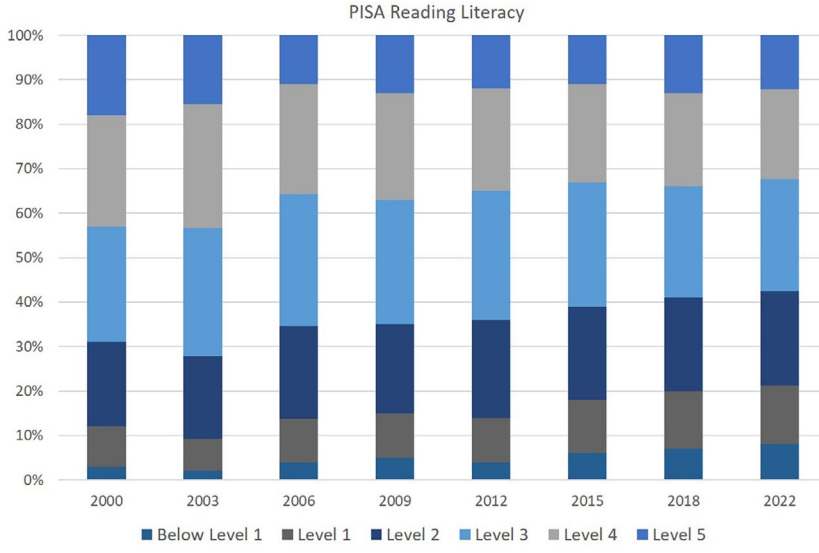

FIGURE 8 Proportions of Australian students meeting each of six proficiency levels in PISA Reading Literacy assessments 2000–2022. N.B. The Level 3 benchmark is the Australian National Proficient Standard.

The mean scores for the PISA Mathematical Literacy test for the representative sample of participating Australian students in all years is shown in Figure 9. The decline in mean scores in Mathematical Literacy is clearly evident in this time series. Australian students had the highest within-country average in 2000 (533 points). This average score steadily declined to a mean of 487 in the latest assessment round in 2022. This change equates to approximately 0.45*SD* on the PISA Mathematical Literacy scale.

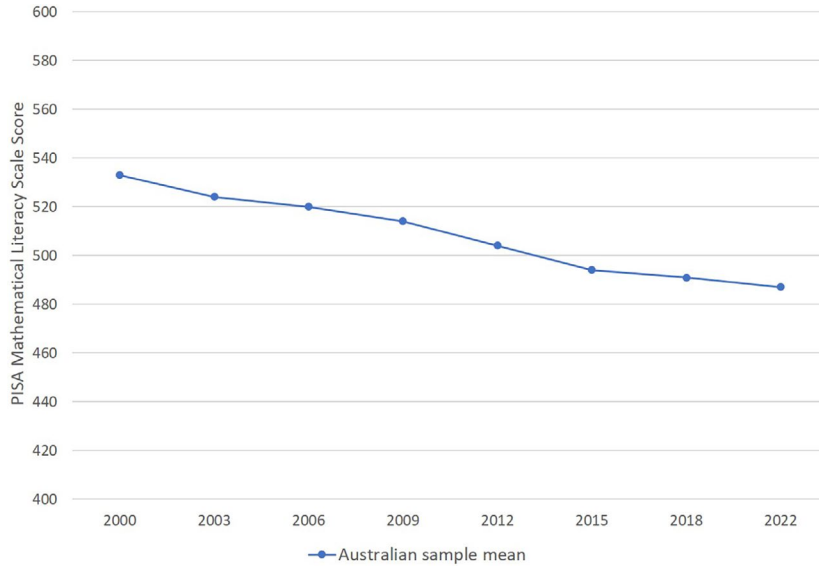

FIGURE 9 Time series showing PISA Mathematical Literacy test mean scores for the representative sample of participating Australian students 2000–2022.

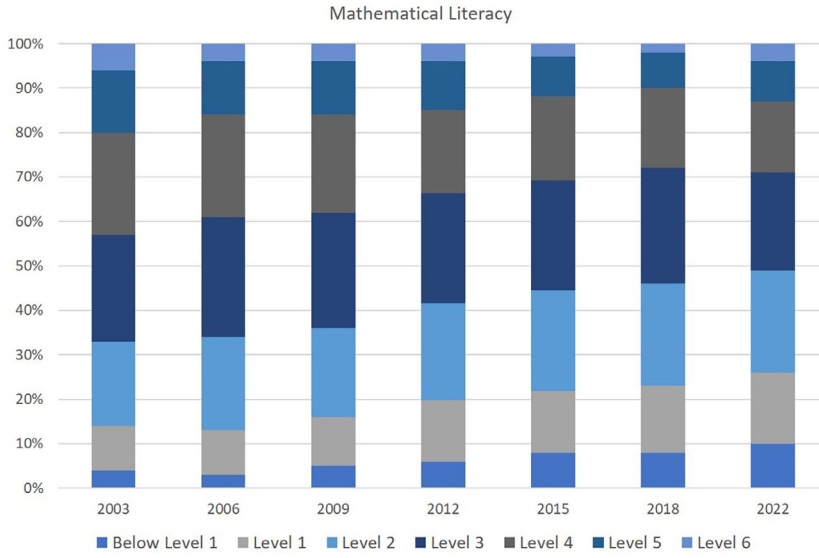

FIGURE 10 Proportions of Australian students meeting each of seven proficiency levels in PISA Mathematical Literacy assessments 2003–2018. N.B. The Level 3 benchmark is the Australian National Proficient Standard.

Figure 10 shows the percentages of students falling into each of seven proficiency levels from 2003 to 2022. Again, similar to the distributions of students in the Reading tests, Figure 10 shows the steady increase in the percentage of students not meeting the Australian national proficient standard (Level 3). In particular, the percentage of students falling into the two lowest proficiency levels increased from 14 per cent in 2003 to 26 per cent in 2022. This trend was

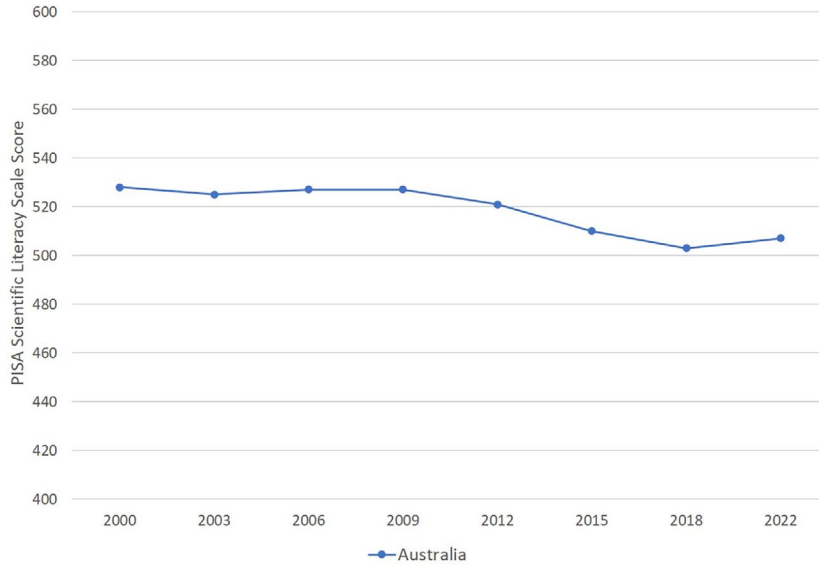

FIGURE 11 Time series showing PISA Scientific Literacy test mean scores for the representative sample of participating Australian students 2000–2022.

accompanied by a decline in the proportion of students in the highest two proficiency bands. In 2003, 20 per cent of Australian students achieved in these highest two bands; in 2022, this had declined to 13 per cent of students, with only 4 per cent achieving in the highest band.

Figure 11 shows the mean scores for Australian students on the PISA Scientific Literacy tests from 2000 to 2022. Similar to both the Reading and Mathematics tests, mean scores declined between 2000 and 2022, though this trend is particularly notable between 2009 and 2018. The change from a mean of 528 points (2000) to 503 points (2018) is approximately equivalent to a change of 0.25*SD*. The difference between 2018 and 2022 (an increase of 4 points to 507) was not statistically significant. Also similar to the other two domains, Figure 12 shows an increase in the number of students in the bottom two proficiency levels for Scientific Literacy, increasing from 13 per cent of students in 2000 to 19 per cent of students falling into the bottom two categories in 2022. This trend is also accompanied by a steady decline of students attaining the top two proficiency standards, falling from 15 per cent in 2000 to 10 per cent in 2018, though 13 per cent of students achieved in the top two categories in the 2022 tests.

3.4 | NAPLAN

In this section, I present data for each of the five domains of NAPLAN: the Literacy domain, which includes Reading, Spelling, Grammar and Writing, and the single Numeracy assessment. In the interests of parsimony, I present mean scores for all calendar years for all four assessed school years in one figure. Since the standard deviations differ for each year, the *y*-axis is scaled to cover the full possible range of scale scores (i.e., 0–1000). The National Report on Schooling Data Portal (ACARA, n.d.) indicates that the writing test used prior to 2011 (narrative writing) was not on the same reporting scale as the writing tests from 2011 onwards (persuasive writing). Therefore, writing test results prior to 2011 are not able to be compared with tests subsequent to 2011. Writing tests are reported from 2011 onwards. In both 2021 and 2022, the Year 9 participation rate across all NAPLAN domains was less than 90 per cent. ACARA (n.d.) indicates this low participation rate means that estimates do not

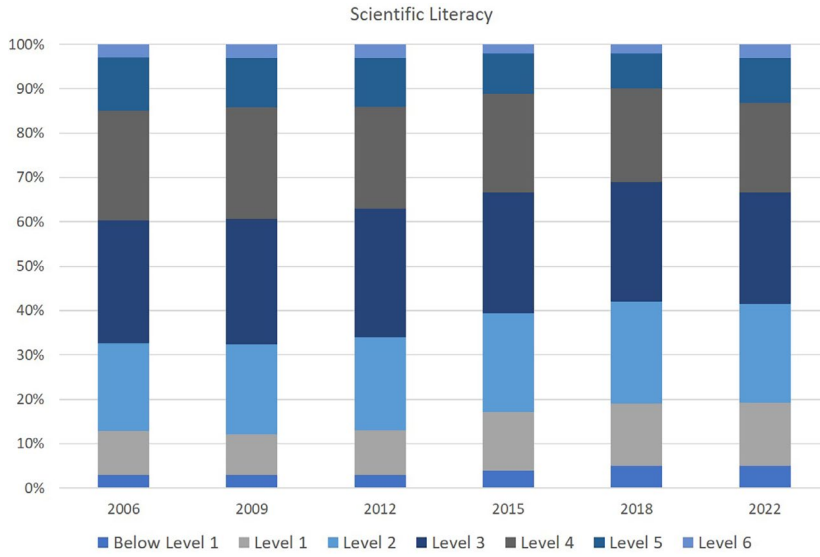

FIGURE 12 Proportions of Australian students meeting each of seven proficiency levels in PISA Scientific Literacy assessments 2006–2022. N.B. The Level 3 benchmark is the Australian National Proficient Standard.

"meet the technical data standards to ensure unbiased results for the calendar year." Finally, scale score changes are transformed into standard deviations (SD) using the SD of the first reported assessment. The SDs differ by year, with generally smaller SDs in the higher years. However, SDs do not change appreciably from cohort to cohort. Unlike the national reporting (ACARA, n.d.), I do not interpret statistical significance of the differences between cohorts given the whole population is sampled for each assessment.

The mean scale scores for the population of Australian students undertaking NAPLAN Reading tests overall years are shown in Figure 13. As for all NAPLAN domains, the mean

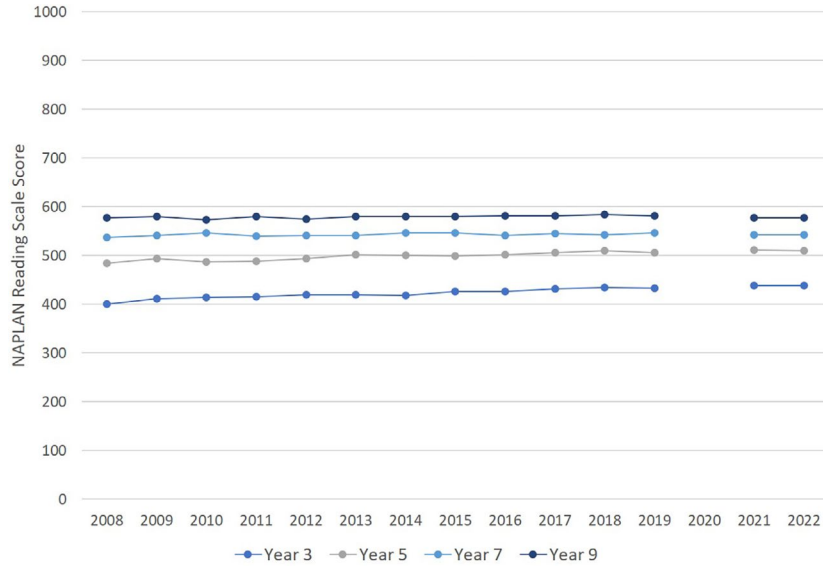

FIGURE 13 Time series showing NAPLAN mean Reading test scores for Years 3, 5, 7 and 9 at all calendar years assessed. *Note*: NAPLAN tests were not undertaken in 2020.

scores increase by school year, with Year 3 represented by the bottom line, increasing through Year 5 and 7 represented by the middle lines and Year 9 mean Reading scores represented by the top line. The Year 3 mean increased steadily from 2008 (400) to 2021/2022 (438). This change can be interpreted as approximately a 0.45*SD* improvement in reading across the Australian population of Year 3 students from 2008 to 2022. The Year 5 mean also increased from 2008 (484) to 2022 (510), an increase of approximately 0.34*SD* over 14 years. By contrast, the Year 7 and Year 9 mean Reading scores remained relatively stable—a trend that was appropriately labelled "flat" by ACARA (n.d.). Year 7 scores ranged from 536 to 546, while Year 9 ranged from 575 to 585. A 10 scale-score difference is equal to approximately 0.15*SD* in Year 7 and Year 9 NAPLAN, though neither year showed notable patterns of improvement or decline.

Mean scores for NAPLAN Spelling (Figure 14) show similar patterns to those for Reading, though with smaller increases in means for Years 3 and 5. Year 3 Spelling increased from 400 in 2008 to 418 in 2022 (+0.23*SD*), while Year 5 increased from 484 to 505 (+0.29*SD*). Year 7 and Year 9 Spelling remained similar in all years' tests. In the Grammar and Punctuation domain (Figure 15), Year 3 was the only year to show increases in mean scores from 2008 (403) to 2022 (433), an improvement of approximately 0.34*SD*. The remaining year levels demonstrated notably stable patterns over the 14 years of tests, with some years returning slightly higher, and other years returning slightly lower average scores.

NAPLAN Numeracy test mean scores for all school years and calendar years are shown in Figure 16. Results in Numeracy for all years have been largely stable since 2008. The most notable exception is the increase in mean scores for Year 5 from 475 in 2008 to a high of 495 in 2021, an improvement of 0.29*SD*. In the remaining year levels, mean scores improved or declined slightly from year to year with no striking overall pattern (i.e., a "flat" trend; ACARA, n.d.).

NAPLAN Writing test means from 2011 to 2022 are in Figure 17. Year 3 Writing means remained stable from 2011 to 2022 with slight increases and declines in each year. The remaining three year levels show a trend that ACARA (n.d.) labels as "u-shaped." That is, writing scores declined slightly from 2011 and then improved, to arrive at a similar mean by 2022.

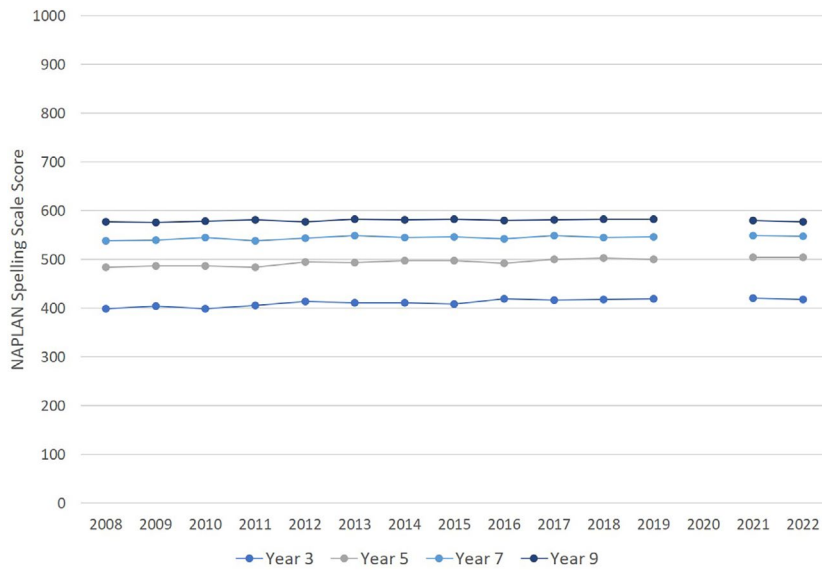

FIGURE 14 Time series showing NAPLAN mean spelling test scores for Years 3, 5, 7 and 9 at all calendar years assessed. *Note*: NAPLAN tests were not undertaken in 2020.

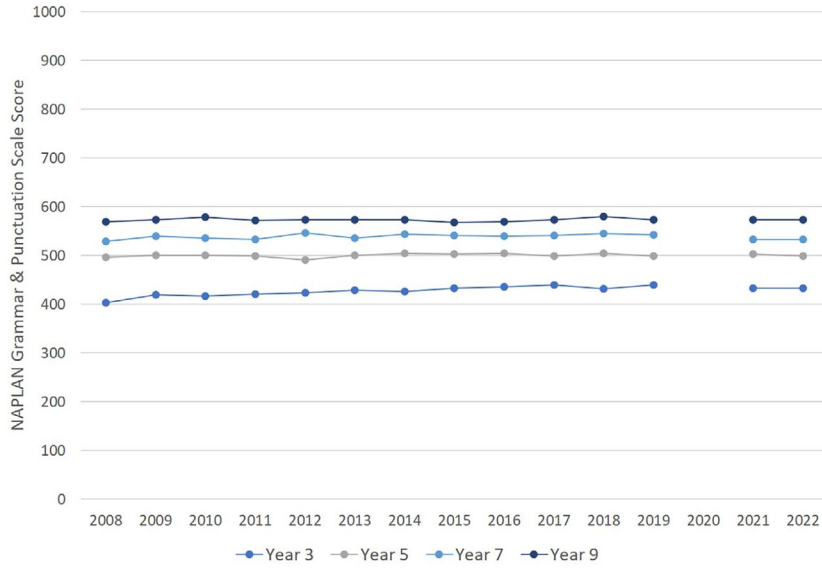

FIGURE 15 Time series showing NAPLAN mean grammar and punctuation test scores for Years 3, 5, 7 and 9 at all calendar years assessed. *Note*: NAPLAN tests were not undertaken in 2020.

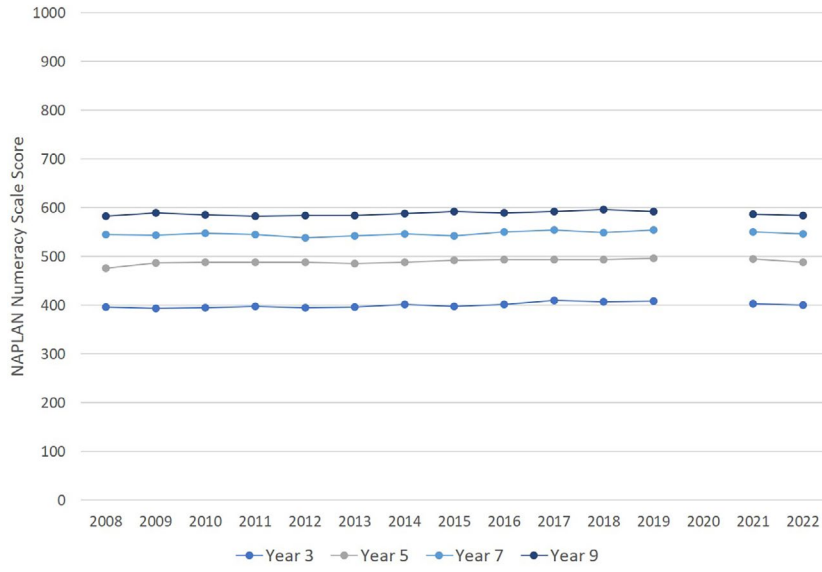

FIGURE 16 Time series showing NAPLAN mean numeracy test scores for Years 3, 5, 7 and 9 at all calendar years assessed. *Note*: NAPLAN tests were not undertaken in 2020.

Figures 18–22 present the proportions of students with scores in each band of the NAPLAN scale for each cohort in each school year. The range of achievement across the 10 possible NAPLAN bands is reported slightly differently for each year level, though the achievement bands are comparable across Years 3–9. For example, Year 3 achievement is reported in six bands: Band 1, which represents those students not meeting the national minimum standard on each NAPLAN test domain, to "at or above" Band 6, which includes

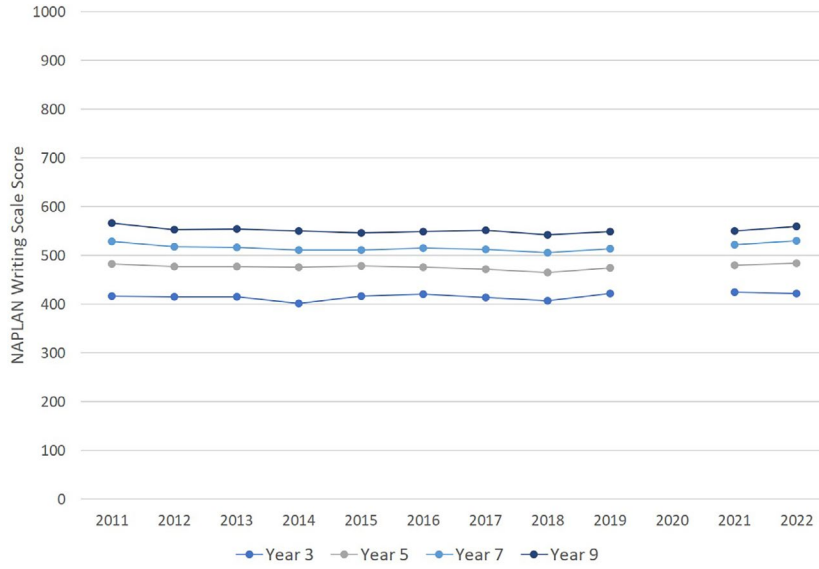

FIGURE 17 Time series showing NAPLAN mean writing test scores for Years 3, 5, 7 and 9 at all calendar years assessed. *Note*: NAPLAN tests were not undertaken in 2020.

FIGURE 18 Proportions of Australian students meeting each of six achievement bands in NAPLAN Reading tests 2008–2022. *Note*: * indicates proportion of students below the national minimum standard band in that year. ** indicates students at or above this band in each year. Bands are colour coded so that the same band has the same colour over the year tests (e.g., Band 6 is green in each quadrant).

students achieving higher than Band 6 in Year 3. In Year 5, the lowest reported group is performing "at or below" Band 3 (i.e., below the national minimum standard for Year 5), while the highest achieving group is at or above Band 8. Students achieving below the national

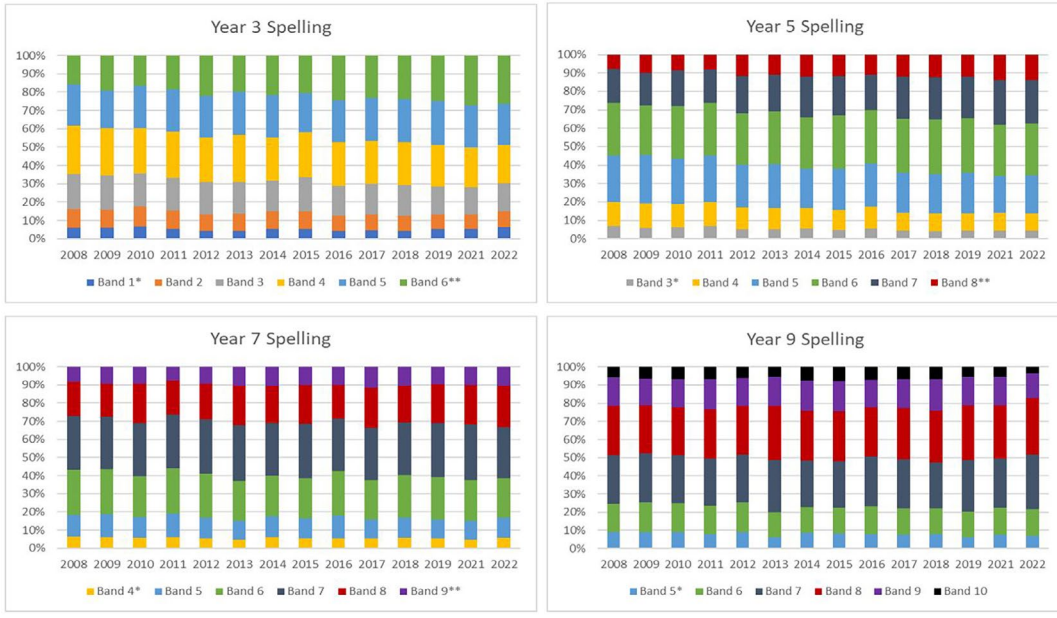

FIGURE 19 Proportions of Australian students meeting each of six achievement bands in NAPLAN Spelling tests 2008–2022. *Note*: * indicates proportion of students below the national minimum standard band in that year. ** indicates students at or above this band in each year. Bands are colour coded so that the same band has the same colour over the year tests (e.g., Band 5 is light blue in each quadrant).

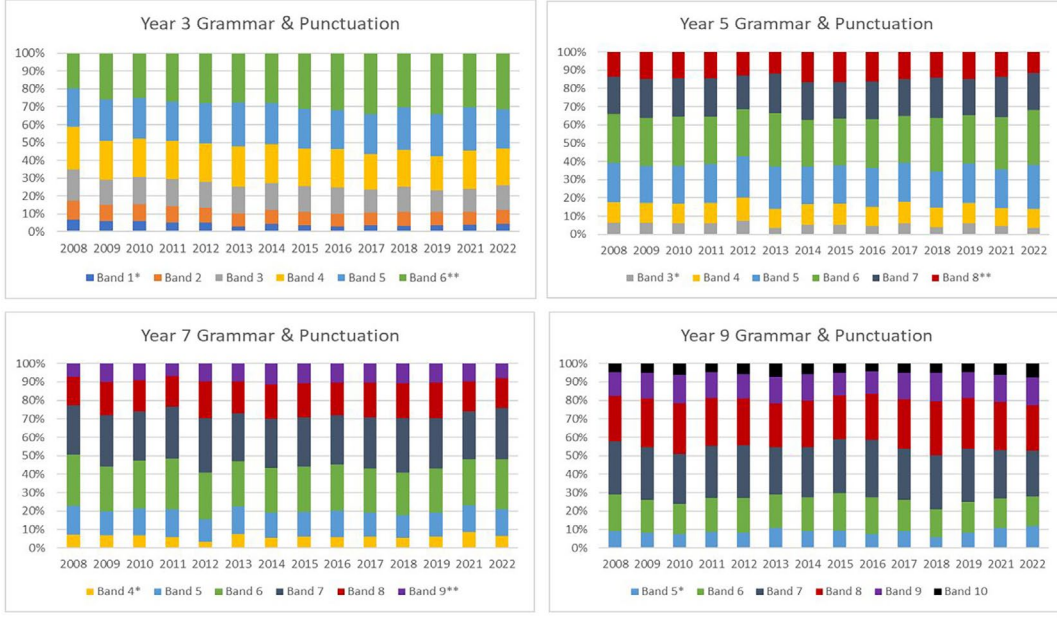

FIGURE 20 Proportions of Australian students meeting each of six achievement bands in NAPLAN Grammar and Punctuation tests 2008–2022. *Note*: * indicates proportion of students below the national minimum standard band in that year. ** indicates students at or above this band in each year. Bands are colour coded so that the same band has the same colour over the year tests (e.g., Band 5 is light blue in each quadrant).

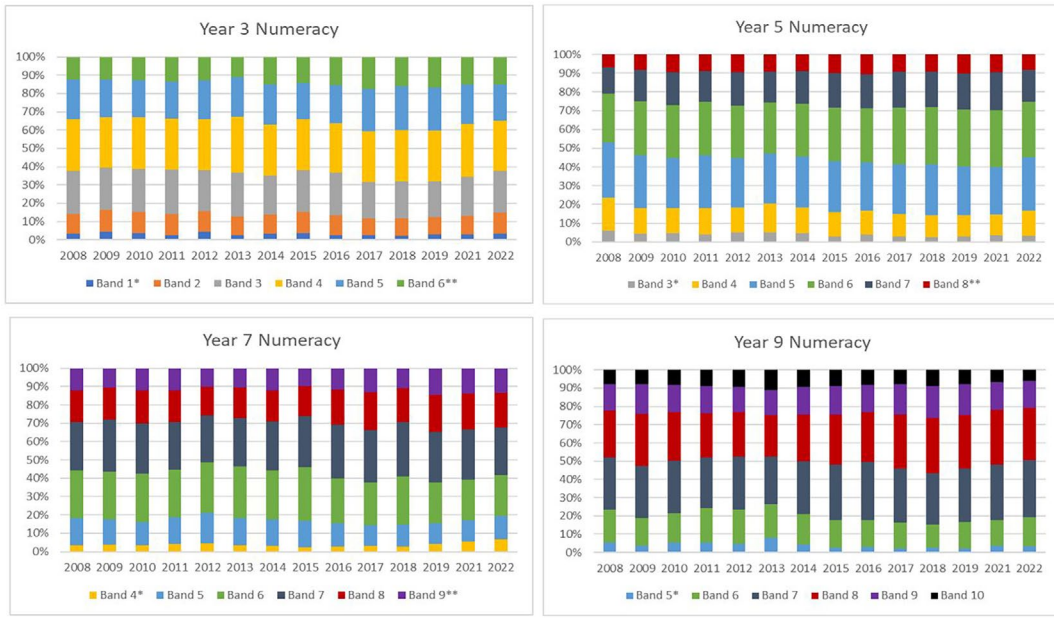

FIGURE 21 Proportions of Australian students meeting each of six achievement bands in NAPLAN Numeracy tests 2008–2022. *Note*: * indicates proportion of students below the national minimum standard band in that year. ** indicates students at or above this band in each year. Bands are colour coded so that the same band has the same colour in each years' tests (e.g., Band 6 is green in each quadrant).

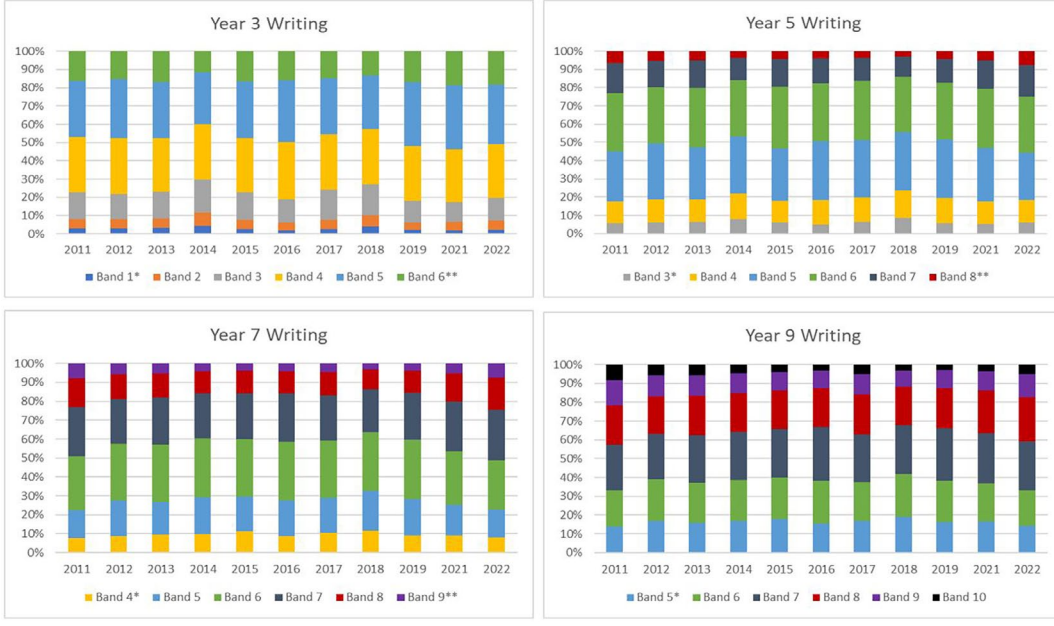

FIGURE 22 Proportions of Australian students meeting each of six achievement bands in NAPLAN Writing tests 2011–2022. * indicates proportion of students below the national minimum standard band in that year. ** indicates students at or above this band in each year. Bands are colour coded so that the same band has the same colour in each years' tests (e.g., Band 6 is green in each quadrant).

minimum standard are grouped in Band 4 in Year 7 and Band 5 in Year 9, while the topmost band is only reported in Year 9 (i.e., Band 10). Figures 18–22 use colours to represent the crossover of each band of achievement as the years progress. That is, Bands 5 and 6 are reported in each year group, and these are coloured light blue and green, respectively. Similarly, Bands 7 and 8 appear in Year 5 and are reported through to Year 9, though in Year 5, they represent the upper end of the achievement distribution, while in Year 9, they represent the middle of the distribution.

For the Reading, Spelling and Grammar tests notable are the increasing proportions of students in Year 3 achieving in Bands 5 and 6 (i.e., the highest achievement bands) from 2008 to 2022. Year 5 Reading and Spelling display a similar pattern, with the percentage of students achieving at Band 7 or 8 increasing by about 10 per cent from 2008 to 2022. Reading also shows small declines in the proportions of students at or below the minimum standard in Years 3 and 5, as does Year 5 Spelling. The proportions of students in the top and bottom bands in the remaining tests for Year 3 and 5, and in all literacy domains for Year 7 and 9, remained fairly stable from 2008 to 2022, with some variation year to year.

The proportion of students in each year level cohort attaining each band in Numeracy and Writing tests also remained relatively stable over the included test years. Notable is the larger proportion of Year 9 students not meeting the minimum standard for Writing across all years 2011 to 2022, compared with Years 3–7. Indeed, the writing domain shows increasing proportions of students not meeting the minimum standard as year-levels progress: less than 5 per cent in, Year 3, increasing to ~5 per cent in Year 5, ~10 per cent in Year 7 and consistently over 10 per cent in Year 9.

It is notable also that declining proportions of students attain the highest bands in all NAPLAN domains as the school years advance. For example, 20–30 per cent of students attain Band 6 in Year 3 Reading, declining to 10–15 per cent at Band 8 in Year 5, around 10 per cent at Band 9 in Year 7 and only 5 per cent attaining Band 10 in Year 9. A similar pattern is evident in all five NAPLAN domains to a greater or lesser extent. Numeracy is an exception, in that just over 10 per cent of students attain the top band in Year 3, 5 and 7 with a smaller proportion attaining the top band in Year 9 (just under 10 per cent).

4 | DISCUSSION

The aim of this paper was to collate information about Australian students' achievement on the four major national and international standardised assessments across all participation years. Standardised assessment data are of central importance in contemporary education policymaking in Australia; however, results from multiple tests are usually reported in isolation, often selectively. Given the dominance of assessment data in school system accountability processes, it is vitally important that data from all current large-scale assessments are easily accessible and interpretable to a broad audience of stakeholders. Collating together and reporting high-level statistics for all educational assessments to date was the main goal of this work.

I selected the four programmes identified in the *Measurement Framework for Schooling in Australia* (ACARA, 2020) since these are intended to be used to evaluate the performance of both students and the education system in Australia. Assessments included the three international programmes, PIRLS, TIMSS and PISA, and one national programme, NAPLAN. The international programmes sample representative groups of Australian students and provide cross-sectional information about each participating cohort that is generalised to the population. NAPLAN is designed to assess the full population of students at four school years (i.e., it is a population census assessment). And provides both cross-sectional and longitudinal information about students' achievement and progress. Evaluating and comparing information

generated by these four assessments using data from all waves provides a more balanced view of Australian students' academic achievement and progress over the last 25 years than considering each in isolation.

Presenting and interpreting year-on-year change in mean scores (or lack thereof), as I have done in this paper, provides only limited information about students' achievement and indeed, no information about variability across the distribution of student achievement. Nonetheless, means are an unbiased measure of central tendency in normally distributed data such as that generated in standardised assessment programmes, and they are the most widely reported statistic. Importantly, information about students' average achievement contributes to public perceptions of the quality of Australian teachers (Mockler, 2022) and influences education policy development (ACARA, 2020). Mean scores are therefore central to a public understanding of Australian students' achievement and progress in school.

Furthermore, changes in mean scores underpin claims about the success (or otherwise) of Australian schools and teachers, and justify calls for policy change (Australian Government, 2023b). A key finding from this paper is the observed differences across different assessments in longitudinal trends of average cohort achievement. For example, both PIRLS (Year 4 students) and Year 3 NAPLAN show immediate improvements in average achievement from the initial rounds, which are subsequently sustained over time. Some Year 5 NAPLAN domains also demonstrate this pattern (e.g., Reading, Spelling and Numeracy), as does the Year 4 TIMSS Mathematics domain. By contrast, Year 5 NAPLAN Numeracy and Writing, and all Year 7 and 9 NAPLAN domains show largely stable patterns of average achievement from 2008 to 2022. Variation is evident from year to year, with higher averages in some years than others, though no systematic process of long-term increase or decline in literacy or numeracy is evident in Years 7 and 9 NAPLAN results. Year 8 TIMSS Maths and Science mean scores also increase or decrease marginally from one round to the next, leading to a largely stable trajectory of cohort averages from 1995 to 2019.

In contrast, PISA Reading, Mathematical and Scientific Literacy tests demonstrate a pattern of steady decline in Australian students' average scores (Figures 7, 9 and 11)—though it is important to note that changes in each test domain from 2015 to 2022 (the three latest PISA rounds) have been small and not statistically significant (De Bortoli et al., 2023). The most pronounced decline was for Mathematical Literacy where the average score declined by an overall ~0.45*SD* from 2000 to 2022. Reading and Scientific Literacy average scores both declined by ~0.25*SD* from 2000 to 2022. These results align with the evaluation of the same assessments by McGaw et al. (2020) and AERO (2023), though I report several additional waves of data for each programme in this study. The results also mirror those of Georgiou (2023) who examined standardised assessments of science in the high school years, including PISA, Year 8 TIMSS, the National Assessment Program for Science Literacy (NAP-SL) and the Validation of Assessment for Learning and Individual Development (VALID) Science assessment undertaken by all public school students in NSW. Georgiou found no evidence of a generalised decline in science achievement across the four assessments, with PISA alone demonstrating consistent declines in population averages.

Examining changes in the proportions of students falling into achievement bands provides an additional (though still limited) piece of information that can inform an understanding of change in mean scores in different assessments. For example, PIRLS Reading tests (Year 4) showed an increase in mean scores from 2011 to 2016, and an increase in students reaching the top band, along with a decrease in the proportion of students in the bottom achievement band. This same pattern is evident in TIMSS Year 4 Mathematics, and NAPLAN Year 3 Reading data. For example, over 14 years, the average improvement in Year 3 students' reading as assessed by NAPLAN was equivalent to 0.45*SD*. This was accompanied by increasing proportions of students in the top band (from 18 per cent in 2008 to 32 per cent in 2022) and decreasing proportions of students falling into the bottom band (from 6 per cent in 2008 to 3 per cent in 2022).

Conversely, all PISA domains demonstrated the reverse pattern: increasing proportions of students falling in the bottom two proficiency levels, and decreasing proportions of students attaining scores in the top level of achievement. For example, in 2000, 18 per cent of Australian students achieved the top level in PISA Reading Literacy tests. This declined to 12 per cent by 2022, accompanied by an increase in students falling in the lowest proficiency level (from 3 per cent in 2000 to 8 per cent in 2022). It is a mathematical fact that if increasing proportions of students score in the bottom tail of a distribution and, simultaneously, declining proportions of students score in the upper tail, the average will decline. We see this effect clearly in the PISA data.

What could be the causes of these differential patterns? There is no easy or straightforward answer to this question. One possible theoretical explanation is suggested by the work of Ruiz-Primo et al. (2002) explaining why educational interventions might demonstrate different effects depending on the assessments used to evaluate them (*c.f.* Slavin & Madden, 2011). Figure 23 provides a visual illustration of the continuum of assessment proximity described by Ruiz-Primo et al. (adapted from their figure on p. 372). The authors argued that the more distant an assessment is from the content of what is learned in the classroom (or the intervention), the smaller the effects that should be observed. Standardised assessments are considered either *distal* or *remote* since they do not assess classroom content in the same depth as more *proximal* assessments, and they are not as sensitive to changes in instruction. Indeed, some educational interventions might not show any effects if assessments are too remote from classroom instruction.

Considering where each of the assessments included in this paper fall on the continuum in Figure 23 can help make sense of the contradictory achievement trends. For example, PISA could be classified as a *remote* assessment because it is not linked to the Australian curriculum and assesses the application of knowledge that may not have been taught to Australian students. While the documentation for PIRLS and TIMSS indicates these assessments are linked to the Australian curriculum, they are still *distal* to classroom learning. NAPLAN could be described as a *distal* assessment in the secondary school years; however, it may fall into the *proximal* assessment category in the primary school years given the documented focus in schools on supporting students to do well in NAPLAN assessments in Years 3 and 5 (Thompson, 2013).

Examining the content and purpose of these different assessments, and evaluating their proximity to the curriculum through the continuum presented in Figure 23, therefore affords

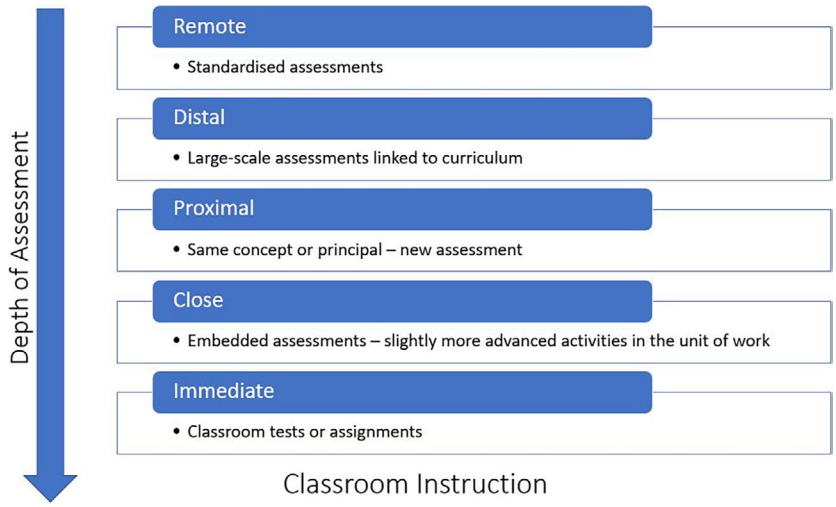

FIGURE 23 Achievement indicators: Instructional sensitivity (adapted from Ruiz-Primo et al., 2002).

some insights into why divergent achievement patterns have arisen. For example, the documentation describing the PISA tests clearly articulates that they are designed to test applied knowledge in each of the included domains—hence the descriptive term "literacy" for each test (Thomson et al., 2019). PISA assessments may be increasingly remote from classroom instruction over time if instruction is more focussed on the basic skills assessed in the remaining three tests, rather than the applied knowledge tested by PISA. By contrast, PIRLS, TIMSS and NAPLAN assess more procedural and technical knowledge in reading, spelling or mathematical operations (i.e., basic skills); these are also more closely aligned with Australian curriculum (ACARA, 2021; Hillman et al., 2023; Thomson et al., 2020), and therefore presumably teachers' classroom practices. Notably, the declines in performance are only observed in the PISA assessments of applied knowledge.

The declining averages in PISA alongside an improving trend in primary school basic skill test results (PIRLS, TIMSS and NAPLAN) has coincided with an increased emphasis on literacy and numeracy in Australian schools, particularly via intensified focus on NAPLAN results (Hardy, 2015). It is possible that curriculum and instructional change efforts at the population level are influencing achievement on these different testing programmes in opposite directions. That is, if increasing amounts of instructional time in schools is devoted to the types of procedural basic skills assessed by NAPLAN, TIMSS and PIRLS, at the expense of time devoted to application, interpretation or academic domain knowledge, then the patterns observed in this study are understandable. That is, PISA becomes more *remote* from classroom instruction while NAPLAN, TIMSS and PIRLS become more *proximal* (Figure 23). Furthermore, the broad stability of average scores in the TIMSS and NAPLAN assessments undertaken in the secondary school years (Years 7–9), alongside the decline in PISA, suggest that if improvement of the population average on both basic skills *and* applied skills is the goal, instruction cannot focus solely on basic skills.

The requirement for generalised "literacy" and "numeracy" instruction in secondary school has gained traction over the last decade, potentially exacerbated by international comparisons arising from PISA (Lingard et al., 2014). Teacher accreditation mandates now require teachers to embed "literacy" and "numeracy" instruction across the curriculum (Australian Institute for Teaching and School Leadership [AITSL], n.d.) in both primary and secondary schools. Interpreting achievement trends with an acknowledgement of the purpose of assessments such as NAPLAN and PISA and their proximity to classroom instruction, provides an explanation for why these kinds of reforms may not have had the impact intended in the secondary school years. Rather than an increasingly restricted focus on "literacy" and "numeracy" as decontextualised skills taught in isolation, it is possible that subject area knowledge is as—if not more—important for improving students' test achievement (Kim & Burkhauser, 2022).

Notwithstanding these observations, justifying causal explanations about average score changes in standardised assessments is extremely difficult. Changes are likely caused by a confluence of factors that are challenging to identify at scale, and individual achievement on these tests is not always an accurate measure of students' true ability (i.e., there is too much error at the individual level; Wu, 2010). Strong causal interpretations about students' attainment on standardised assessments can only be made if there are no alternative explanations for observed data patterns (Shadish et al., 2002). It is clear from the volume of often contradictory commentary and interpretation that appears with the release of data from each new round of assessments, that there are multiple causal interpretations of change in assessment scores, both positive and negative.

Despite the limitations of extrapolating from standardised assessment scores to classroom or teacher practices, and the likelihood that change in classroom practice may not show strong effects in standardised assessments (Ruiz-Primo et al., 2002; Slavin & Madden, 2011), there is a strong contemporary trend toward doing so both in government policy documents and educational think-tank advice (AERO, 2023; Hunter & Parkinson, 2023). As indicated

above, the causative factors at work that contribute to students' level of achievement on these assessments are arguably too complex for direct recommendations to be made about the specific teaching practices that may contribute to improved achievement on these types of standardised tests. Furthermore, average results in these assessments are distal or remote from continuous teaching and assessment practices undertaken in schools (Cumming et al., 2019; Ruiz-Primo et al., 2002): results on standardised assessments are different to in-school achievement or progress, and the two types of data may not always closely match for any individual student (Lee et al., 2019). The assessments that are most closely aligned to classroom instruction, and therefore most amenable to changes in classroom practice, are not standardised tests.

Finally, it is possible that the improvement in average basic skill attainment in the primary school years observed in PIRLS, TIMSS and NAPLAN is due simply to test familiarity (Thompson, 2013). That is, students are now explicitly exposed to standardised assessments from at least Year 3, whereas prior to the advent of NAPLAN (in 2008), this practice was likely not as prevalent across the population. Nonetheless, the Australian federal government is now looking toward initial teacher education (ITE) as a potential target for improvements, with the unfounded theoretical belief that changes to ITE should have flow-on effects to improved student achievement on standardised tests (Australian Government, 2021).

5 | CONCLUSION

There is no strong evidence that the achievement of Australian students has suffered a precipitous decline over the past 25 years, notwithstanding proclamations of media commentators (Kelly, 2021), educational think-tanks (Hunter & Parkinson, 2023), or education ministers (Loughland & Thompson, 2016). Nonetheless, while these analyses show some improvements in average basic skills scores over the last 25 years, achievement "gaps" have remained remarkably persistent even in the context of increasing school accountability made possible by the advent of standardised school assessment programmes (Adams et al., 2020; Cumming et al., 2020). The work of education policymakers should be informed by an accurate, long-term view of the progress of Australian students' achievement—acknowledging both the positives and the areas for potential improvement. Alongside within-school factors, systemic problems, such as inequitable funding models (Chesters, 2018; Connors & McMorrow, 2015) or socioeconomic stratification between schools (Chesters & Daly, 2017; Sciffer et al., 2022), also need to be considered. Without such a view, policy may be doomed to repeat initiatives that have not led to improvements, since an improvement can only be defined with complete and accurate data and a consistent theoretical framework. I hope the insights presented in this paper can therefore contribute to future educational system evaluation and policymaking.

AUTHOR CONTRIBUTION

Sally Larsen: Conceptualization; investigation; writing – original and revised drafts; methodology; formal analysis.

ACKNOWLEDGEMENTS

I would like to acknowledge the thoughtful comments of William Coventry and Genevieve Thraves on the development of this paper. Open access publishing facilitated by University of New England, as part of the Wiley - University of New England agreement via the Council of Australian University Librarians.

ORCID

Sally A. Larsen https://orcid.org/0000-0001-5742-8444

REFERENCES

Adams, E.K., Hancock, K.J. & Taylor, C.L. (2020) Student achievement against national minimum standards for reading and numeracy in years 3, 5, 7 and 9: a regression discontinuity analysis. *Australian Journal of Social Issues*, 55(3), 275–301. Available from: https://doi.org/10.1002/ajs4.124

Ainley, J., Cloney, D. & Thompson, J. (2020) Does student grade contribute to the declining trend in programme for international student assessment reading and mathematics in Australia? *Australian Journal of Education*, 64(3), 205–226. Available from: https://doi.org/10.1177/0004944120948654

Anderson, R. & Anderson, C. (2020) Grade repetition and boys' risk of being repeated in early schooling in Queensland, Australia. *Journal of Psychologists and Counsellors in Schools*, 30(2), 146–158. Available from: https://doi.org/10.1017/jgc.2019.5

Australian Curriculum, Assessment and Reporting Authority. (2020) *Measurement Framework for Schooling in Australia 2020*. ACARA. Available from: https://www.acara.edu.au/reporting/measurement-framework-for-schooling-in-australia

Australian Curriculum, Assessment and Reporting Authority. (2021) *National Assessment Program—Literacy and Numeracy*: *National Report for 2021*. ACARA. Available from: https://www.nap.edu.au/naplan/results-and-reports

Australian Curriculum Assessment and Reporting Authority. (2022) *National Assessment Program – Literacy and Numeracy 2021*: *Technical Report*. ACARA. Available from: https://www.nap.edu.au/naplan/results-and-reports

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (2023) NAPLAN national results. Available from: https://www.acara.edu.au/reporting/national-report-on-schooling-in-australia/naplan-national-results

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (n.d.) *NAPLAN National Report*. Available from https://www.acara.edu.au/reporting/national-report-on-schooling-in-australia/national-report-on-schooling-in-australia-data-portal/naplan-national-report [12th May 2023].

Australian Education Research Organisation [AERO]. (2023) *Benchmarking performance*: *Future directions for Australia's National Assessment Program*. AERO. Available from: https://www.edresearch.edu.au/resources/benchmarking-performance-future-directions-australias-national-assessment-program

Australian Government. (2021) *Next steps*: *Report of the quality initial teacher education review*. Available from: https://www.education.gov.au/quality-initial-teacher-education-review/resources/next-steps-report-quality-initial-teacher-education-review

Australian Government. (2023a) *National Assessment Program—Programme for International Student Assessment*. Available from: https://www.education.gov.au/national-assessment-program/national-assessment-program-programme-international-student-assessment

Australian Government. (2023b) *Review to inform a better and fairer education system*: *consultation paper*. Available from: https://www.education.gov.au/review-inform-better-and-fairer-education-system/resources/better-and-fairer-education-system-consultation-paper

Australian Institute for Teaching and School Leadership [AITSL] (n.d.) Australian Professional Standards for Teachers. Available from https://www.aitsl.edu.au/standards

Ball, S.J. (2015) Education, governance and the tyranny of numbers. *Journal of Education Policy*, 30(3), 299–301. Available from: https://doi.org/10.1080/02680939.2015.1013271

Briggs, D.C. & Weeks, J.P. (2009) The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3–14. Available from: https://doi.org/10.1111/j.1745-3992.2009.00158.x

Chesters, J. (2018) The marketisation of education in Australia: does investment in private schooling improve post-school outcomes? *Australian Journal of Social Issues*, 53(2), 139–157. Available from: https://doi.org/10.1002/ajs4.38

Chesters, J. & Daly, A. (2017) Do peer effects mediate the association between family socio-economic status and educational achievement? *Australian Journal of Social Issues*, 52(1), 63–77. Available from: https://doi.org/10.1002/ajs4.3

Connors, L. & McMorrow, J. (2015) *Imperatives in schools funding: equity, sustainability and achievement*. Melbourne, Australia: ACER Press. https://research.acer.edu.au/aer/14

Cowger, C.D. (1984) Statistical significance tests: scientific ritualism or scientific method? *Social Service Review*, 58(3), 358–372.

Cumming, J., Goldstein, H. & Hand, K. (2020) Enhanced use of educational accountability data to monitor educational progress of Australian students with focus on Indigenous students. *Educational Assessment, Evaluation and Accountability*, 32(1), 29–51. Available from: https://doi.org/10.1007/s11092-019-09310-x

Cumming, J.J., Van Der Kleij, F.M. & Adie, L. (2019) Contesting educational assessment policies in Australia. *Journal of Education Policy*, 34(6), 836–857. Available from: https://doi.org/10.1080/02680939.2019.1608375

De Bortoli, L., Underwood, C. & Thomson, S. (2023) *PISA 2022. Reporting Australia's results. Volume I: Student performance and equity in education*. Australian Council for Educational Research. Available from: https://doi.org/10.37517/978-1-74286-725-0

Department of Education, Skills and Employment (Ed.) [DESE]. (2019) *Alice Springs (Mparntwe) Education Declaration* [Electronic resource]. DESE. Available from: https://www.dese.gov.au/alice-springs-mparntwe-education-declaration/resources/alice-springs-mparntwe-education-declaration

Georgiou, H. (2023) Are we really falling behind? Comparing key indicators across international and local standardised tests for Australian high school science. *Research in Science Education*, 53, 1205–1220. Available from: https://doi.org/10.1007/s11165-023-10129-2

Gillis, S., Polesel, J. & Wu, M. (2016) PISA data: raising concerns with its use in policy settings. *The Australian Educational Researcher*, 43(1), 131–146. Available from: https://doi.org/10.1007/s13384-015-0183-2

Gonzales, E.J. & Foy, P. (1997) Estimation of sampling variability, design effects, and effective sample sizes. In: Martin, M.O. & Kelly, D.L. (Eds.) *Third international mathematics and science study: technical report, volume 2, implementation and analaysis*. Chestnut Hill: Center for the Study of Testing, Evaluation, and Educational Policy. Boston College. Available from: https://timssandpirls.bc.edu/timss1995i/TIMSSPDF/TR2book.pdf

Hardy, I. (2015) A logic of enumeration: the nature and effects of national literacy and numeracy testing in Australia. *Journal of Education Policy*, 30(3), 335–362. Available from: https://doi.org/10.1080/02680939.2014.945964

Hill, C.J., Bloom, H.S., Black, A.R. & Lipsey, M.W. (2008) Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172–177. Available from: https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hillman, K., O'Grady, E., Rodrigues, S., Schmid, M. & Thomson, S. (2023) *Progress in Reading Literacy Study: Australia's results from PIRLS 2021*. Australian Council for Education Research. Available from: https://doi.org/10.37517/978-1-74286-693-2

Hunter, J. & Parkinson, N. (2023) *The new NAPLAN results are a wake-up call*. Grattan Institute. Available from: https://grattan.edu.au/news/the-new-naplan-results-are-a-wake-up-call/

Kelly, P. (2021) Lessons in failure on education need to be learnt. *The Australian*. Available from: https://www.theaustralian.com.au/inquirer/lessons-in-failure-on-education-need-to-be-learnt/news-story/caf28295dab28deffb81758a65411997

Kim, J.S. & Burkhauser, M.A. (2022) Teaching for transfer can help young children read for understanding. *Phi Delta Kappan*, 103(8), 20–24. Available from: https://doi.org/10.1177/00317217221100006

Kraft, M.A. (2020) Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. Available from: https://doi.org/10.3102/0013189X20912798

Larsen, S.A., Little, C.W. & Coventry, W.L. (2021) Exploring the associations between delayed school entry and achievement in primary and secondary school. *Child Development*, 92(2), 774–792. Available from: https://doi.org/10.1111/cdev.13440

Lee, J., McArthur, W. & Ellis, N.J. (2019) NAPLAN versus in-school assessment: how similar or different are students' results? *Curriculum and Teaching*, 34(2), 5–25. Available from: https://doi.org/10.7459/ct/34.2.02

Lingard, B. (2011) Policy as numbers: accounting for educational research. *The Australian Educational Researcher*, 38(4), 355–382. Available from: https://doi.org/10.1007/s13384-011-0041-9

Lingard, B. & Sellar, S. (2013) 'Catalyst data': perverse systemic effects of audit and accountability in Australian schooling. *Journal of Education Policy*, 28(5), 634–656. Available from: https://doi.org/10.1080/02680939.2012.758815

Lingard, B., Sellar, S. & Savage, G.C. (2014) Re-articulating social justice as equity in schooling policy: the effects of testing and data infrastructures. *British Journal of Sociology of Education*, 35(5), 710–730. Available from: https://doi.org/10.1080/01425692.2014.919846

Loughland, T. & Thompson, G. (2016) The problem of simplification: think-tanks, recipes, equity and 'Turning around low-performing schools'. *The Australian Educational Researcher*, 43(1), 111–129. Available from: https://doi.org/10.1007/s13384-015-0190-3

McGaw, B., Louden, W. & Wyatt-Smith, C. (2020) *NAPLAN Review: Final Report*. State of NSW, State of Queensland, State of Victoria and Australian Capital Territory. Available from: https://naplanreview.com.au/

Mockler, N. (2022) *Constructing teacher identities: how the print media define and represent teachers and their work*. Sydney, Australia: Bloomsbury Publishing.

OECD. (2015) *Universal basic skills: what countries stand to gain*. Paris, France: OECD Publishing. Available from: https://doi.org/10.1787/9789264234833-en

Paris, S.G. (2005) Reinterpreting the development of reading skills. *Reading Research Quarterly*, 40(2), 184–202. https://doi.org/10.1598/rrq.40.2.3

Ruiz-Primo, M.A., Shavelson, R.J., Hamilton, L. & Klein, S. (2002) On the evaluation of systemic science education reform: searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39(5), 369–393. Available from: https://doi.org/10.1002/tea.10027

Savage, G.C., Sellar, S. & Gorur, R. (2013) Equity and marketisation: emerging policies and practices in Australian education. *Discourse: Studies in the Cultural Politics of Education*, 34(2), 161–169. Available from: https://doi.org/10.1080/01596306.2013.770244

Sciffer, M.G., Perry, L.B. & McConney, A. (2022) The substantiveness of socioeconomic school compositional effects in Australia: measurement error and the relationship with academic composition. *Large-Scale Assessments in Education*, 10(1), 21. Available from: https://doi.org/10.1186/s40536-022-00142-8

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Boston, USA: Houghton Mifflin Co.

Slavin, R. & Madden, N.A. (2011) Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, 4(4), 370–380. Available from: https://doi.org/10.1080/19345747.2011.558986

Thompson, G. (2013) NAPLAN, MySchool and Accountability: teacher perceptions of the effects of testing. *International Education Journal: Comparative Perspectives*, 12(2), 2. Available from: https://openjournals.library.sydney.edu.au/IEJ/article/view/7456

Thomson, S., Bortoli, L.D., Underwood, C. & Schmid, M. (2019) *PISA 2018: reporting Australia's results. Volume I student performance*. Camberwell, Australia: OECD Programme for International Student Assessment (PISA) Australia. Available from: https://research.acer.edu.au/ozpisa/35

Thomson, S., Wernert, N., Rodrigues, S. & O'Grady, E. (2020) *TIMSS 2019 Australia*: *Volume 1—Student Performance*. Australian Council for Education Research. Available from: https://doi.org/10.37517/978-1-74286-614-7

Verger, A., Fontdevila, C. & Parcerisa, L. (2019) Reforming governance through policy instruments: how and to what extent standards, tests and accountability in education spread worldwide. *Discourse: Studies in the Cultural Politics of Education*, 40(2), 248–270. Available from: https://doi.org/10.1080/01596306.2019.1569882

Verger, A., Parcerisa, L. & Fontdevila, C. (2019) The growth and spread of large-scale assessments and test-based accountabilities: a political sociology of global education reforms. *Educational Review*, 71(1), 5–30. Available from: https://doi.org/10.1080/00131911.2019.1522045

Wiliam, D. (2019) Some reflections on the role of evidence in improving education. *Educational Research and Evaluation*, 25(1–2), 127–139. Available from: https://doi.org/10.1080/13803611.2019.1617993

Wu, M. (2010) Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, 29(4), 15–27. Available from: https://doi.org/10.1111/j.1745-3992.2010.00190.x

Wu, M. & Hornsby, D. (2014) Inappropriate uses of NAPLAN results. *Practically Primary*, 19(2), 16–17. Available from: https://doi.org/10.3316/informit.320310592534323

AUTHOR BIOGRAPHY

Dr **Sally A. Larsen** is a senior lecturer in the School of Education at the University of New England. She taught English in secondary schools before completing a PhD examining developmental patterns in National Assessment Program: Literacy and Numeracy data. Dr **Larsen** is interested in applications of quantitative methods in education research, in particular, longitudinal structural equation modelling and multilevel modelling. Her research investigates variability in patterns of growth in reading and maths, and predictors of development in these skills across the school years.

How to cite this article: Larsen, S.A. (2025) Are Australian students' academic skills declining? Interrogating 25 years of national and international standardised assessment data. *Australian Journal of Social Issues*, 60, 302–333. Available from: https://doi.org/10.1002/ajs4.341