Research in Science Education (2023) 53:1205–1220
https://doi.org/10.1007/s11165-023-10129-2

# Are We Really Falling Behind? Comparing Key Indicators Across International and Local Standardised Tests for Australian High School Science

**Helen Georgiou**[1]

Accepted: 27 August 2023 / Published online: 19 September 2023


**Abstract**
There has been a strong narrative in Australia of falling attainment in high school science, with much of the campaign informed by results from international standardised tests such as Programme for International Student Assessment (PISA), which shows a year-on-year decline in scientific literacy of Australian 15-year-old students. These results have been used to justify significant policy and curriculum reform, despite the known limitations of PISA and a lack of additional evidence to support this decline in other tests. In this paper, results from standardised tests administered in Australia will be compared to create a fulsome picture of attainment for high school science students. Reports include both the compilation of data from existing reports and new analyses. With the latest (2018/9) reports from PISA, Trends in International Mathematics and Science Study (TIMSS), and National Assessment Program for Scientific Literacy (NAP-SL) (an Australian test of Science Literacy) and data shared by the NSW Department of Education on 'The Validation of Assessment for Learning and Individual Development' (VALID) test for the years 2015, 2016, 2017, and 2018, this offers the most complete picture of student attainment in science to date. Results show that there are disagreements between tests on cohort achievement over time and distribution of attainment at different 'proficiency levels'. These results suggest caution when using these key results from these tests to inform policy and pedagogy.

**Keywords** PISA · TIMSS · High school science · Assessment

## Introduction

Educational policy, both nationally and internationally, has been focused on the creation of a Science, Technology, Engineering, and Mathematics (STEM)–focused workforce and scientifically literate citizenry (GBC and Education Commission, 2019: PwC, 2015). With the implementation of standardised tests on a global scale, many countries are discovering that their students are not achieving at the levels that were

✉ Helen Georgiou
helengeo@uow.edu.au

[1] University of Wollongong, Wollongong, NSW 2522, Australia

anticipated or required to meet these objectives (e.g. Haugsbakk, 2013; Waldow, 2009). In Australia, over the last decade, concerns around the decline in achievement of students in international standardised tests have heightened, with sliding national rankings increasingly lamented across academe, politics, and society (e.g. Hopfenback, 2018; Hildebrand, 2023).

Consistent with its broad international impact on education policy (e.g. Froese-Germain, 2010; Pons, 2017; Wiseman, 2013), the Programme for International Student Assessment (PISA) is quoted as a central source of information in several key documents in Australia, including position papers from the Office of the Chief Scientist (e.g. Office of the Chief Scientist, 2014), the Education Council's National STEM School Education Strategy (Education Council, 2015), and documents outlining the justification for the New South Wales (NSW) Standards Authority's curriculum transformation (NESA, 2020). Concerns about falling attainment pervade political speeches and inspire a near constant stream of media articles with headlines such as 'our system is failing' (Hildebrand, 2023) and 'why Australia's students keep falling behind' (Hare, 2022).

However, the assumption that there is a real and significant decline in Australian students' achievement in science has not been critically explored. Most concerning is that PISA is almost exclusively used to establish the idea that Australian students are getting worse in science, particularly given that the limitations of PISA have been known for some time (e.g. Zhao, 2020). Amongst those most relevant here include that PISA is not a curriculum-based test and that PISA results are not consistent with other standardised tests. Trends in International Mathematics and Science Study (TIMSS), in fact, shows Australia's high school science performance improving, both in terms of rank (against other countries) and in absolute terms, over time. In Australia, some students also sit a national curriculum-based standardised test, the National Assessment Program for Scientific Literacy (NAP-SL), and in the state of NSW, the largest educational jurisdiction in the country, all year, 8 students in public schools are required to take the Validation of Assessment for Learning and Individual Development (VALID). PISA, TIMSS, and NAP-SL only test a sample of students and are run every few years, and VALID results are not presented to the public in reports in the same way as these other tests. As such, we have both an incomplete picture and a contradictory one. In this paper, we compare the findings across the two international and one national test and undertake an analysis of the state-based VALID data to come to an understanding of what we should and should not be concerned about when it comes to student achievement in science. The focus of this paper is students at the high school level only, and, as such, data available for students at the primary levels are not included.

## Background

### International and National Standardised Tests

Each individual international or national standardised test has a different purpose, design, and administration. In Australia, we draw on PISA, TIMSS, and, to a lesser extent, our national standardised test, NAP-SL. VALID is used only in the state of New South Wales (NSW). An overview of these tests is provided below and summarised in Table 1.

**Table 1** An overview of standardised tests of science administered to Australian high school students

| Test | Admin. body | Jurisd. | Focus | Stage | Frequency | Last admin | Sample | Key figures |
|---|---|---|---|---|---|---|---|---|
| PISA | OECD | Intl | Science, mathematics, reading, other | 15 years old (mainly years 9 and 10) | 3-year cycle (science focus every 9 years, reading main focus in 2018) | 2018* | Random stratified | 79 countries<br>4000–8000 students per country |
| TIMSS | IEA | Intl | Mathematics, science | Year 4, year 8 | 4-year cycle | 2019 | Random stratified | 64 countries<br>14,950 Australian students |
| NAP-SL | ACARA | Aus | Science literacy | Year 6 (year 10 in 2018) | 3-year cycle | 2018* | Random stratified | 5578 (yr 6)<br>3043 (yr 10) |
| VALID | DOE NSW | NSW | Science | Year 8 | Yearly | 2020* | Full population (public schools) | ~ 45,000 students |

*All due for 2022 administrations after COVID disruptions in 2020 and 2021. Reports from PISA, TIMSS, and NAP-SL are usually prepared in the year following administration. VALID data are not published in a public report

*OECD* Organisation for Economic Co-operation and Development; *IEA* International Association for the Evaluation of Educational Achievement; *ACARA* Australian Curriculum, Assessment and Reporting Authority; *DOE* NSW Department of Education

### PISA

The OECD's PISA evaluates 15-year-old students' ability to apply knowledge and skills in three different domains, science, mathematics, and reading (additional tests have been trialled in recent years, including financial literacy and creativity). The science assessment 'focuses on measuring students' ability to engage with science-related issues and with the ideas of science, as reflective citizens' (OECD, 2019, p.112) and as such, PISA's assessment of 'Scientific Literacy' essentially represents a students' ability to apply their scientific knowledge to everyday situations. PISA takes the form of a 2-hour computer-based cognitive assessment and consists of short answer and multiple-choice questions in response to a stimulus. PISA is administered every 3 years, with only one domain emphasised each time, though this structure is changing to include more of each domain in non-focus years. Science was the major domain in 2006 and 2015.

### TIMSS

TIMSS evaluates year 4 and year 8 students' mathematics and science achievement, again, on a 4-year cycle. TIMSS is a project of the IEA (International Association for the Evaluation of Educational Achievement), and in Australia, its administration and reporting are managed by the Australian Council for Educational Research (ACER). The focus for TIMSS is 'factual and procedural knowledge' in mathematics and science as represented in a defined curriculum (TIMSS surveys representatives from each country to confirm curriculum content to be assessed). From 2019, TIMSS has been administered online and is 102 min in duration in year 4 and 122 min for year 8.

For both PISA and TIMSS, approximately half of the items are multiple-choice or other closed formats (such as drag and drop or complex multiple choice), with the other half being constructed response (which can range from a word or value to an extended response). Students' attainment is measured in terms of an overall score and level of proficiency, with both tests establishing level 3 as the benchmark for their respective populations. Average scores, such as those reported for states, nations, and different sample groups, are not 'raw' scores but through 'plausible values' which are essentially theory-informed estimates.

### NAP-SL

The NAP-SL tests Australian year 6 and 10 (from 2018) students' science literacy. Science literacy is defined as students' ability to 'use scientific knowledge, understanding, and inquiry skills to identify questions, acquire new knowledge, explain science phenomena, solve problems and draw evidence-based conclusions in making sense of the world, and to recognise how understandings of the nature, development, use and influence of science help us make responsible decisions and shape our interpretations of information' (ACARA, 2019). ACARA claims that this definition is consistent with the PISA definition of scientific literacy. NAP-SL is a national assessment program, administered by the Australian Curriculum, Assessment, and Reporting Authority (ACARA) every 3 years. NAP-SL is based on the national curriculum and thus assesses three strands (science understanding, science and a human endeavour, and science inquiry skills). NAP-SL 2018 was an online

test. The 'proficient' standards are 'the boundary between levels 2 and 3' for year 6 and 'the boundary between 3 and 4' for year 10.

### VALID

The VALID is a state (NSW) test, compulsory for year 8 students in public schools in NSW and optional for year 6 and 10 students and students in independent schools. VALID is a yearly, curriculum-based test, assessing specific outcomes from the NSW Science Syllabus (https://curriculum.nsw.edu.au/learning-areas/science). VALID is an online tests and consists of a range of item types, with three extended response questions.

All abovementioned tests also include a contextual survey, not reported on here, which asks students, teachers, and school leaders to comment on a range of non-academic factors, such as well-being and classroom environment.

Most International large-scale assessments (ILSAs), such as PISA or TIMSS, were designed to improve educational outcomes through policy changes resulting from the comparison between different countries. For instance, the OECD explains of PISA that observing the similarities and differences between educational policies and practices 'enable(s) researchers and others to observe what is possible for students to achieve and what environment is more likely to facilitate their learning' (Thomson et al., 2019, p xiv). The IEA also explains that 'the goal of TIMSS is to provide comparative information about educational achievement across countries in order to improve teaching and learning in mathematics and science' (Thomson et al., 2020 p. xiii). The aims of ILSAs are thus very different from other standardised tests, which provide different utility depending on the design and nature of the test. VALID, for example, was designed as a formative assessment tool, which provides feedback to students, parents, and teachers on students' achievement and allows for local changes/improvements to be made in response to student difficulties. NAPLAN, the literacy and numeracy-based 'cousin' of NAP-SL, whether intentional or not, is also used to measure individual school performance and progress with results publicly available on the 'MySchool' website. Despite these significant differences, each test purports to measure some aspect of student achievement in science.

Compared to all other ILSAs and more local standardised tests, PISA has received significantly more attention, both in the literature and more generally. For instance, Hopfenbeck et al. (2018) show that Australia has the second highest number of published articles of PISA, second only to the USA, and Jerrim (2023) reports that, on average, PISA results receive around 10 times more attention than TIMSS on search engines. In Australia, 80% of all ILSA searches relate to PISA (compared with 17% for TIMSS and 3% for PIRLS). The reasons for this are regularly under conjecture (and criticism), though it is generally understood that the OECD aspired that PISA would provide a point of departure from existing ILSAs such as TIMSS by testing students' *application* of knowledge to new situations, rather than simply their reproduction of scientific knowledge or scientific facts (Zhao, 2020). This outcome is seen to provide a much-desired economic association, as it is assumed that it is a proxy for 'work readiness' and even linked to potential changes in GDP (Araujo, 2017).

All tests mentioned in this report are carefully designed in accordance with item response theory (IRT) to ensure computed scores tell us something about what each test claims to measure. Available technical reports from PISA, TIMSS, and NAP-SL provide information around how the validity, reliability, and accuracy of tests are assured, including information around how items were developed, how sampling was managed,

and how test reliability was assured. In interpreting the meaning of scores and how they change over time, validity concerns become significant. Validity refers to the degree to which the test provides an accurate measure of what it states it is measuring. Validity is known to be an incredible challenge in test design (Cizek 2012; Kane, 2016). Crucially, validity is considered to apply to inferences about test scores, not the tests themselves (Kane, 2016). For example, whilst scores themselves might be valid in terms of measuring a distinct construct, whether this construct can be understood as 'scientific literacy' or 'problem solving' (and what this means) is much less assured (Hopfenbeck et al., 2016; Liou and Bulut, 2020; Rindermann and Baumeister, 2015; Wu, 2009). This is problematic, particularly for PISA, which identifies an overall construct (readiness for 'real world' problem solving), which is more ambiguous and multifaceted than curriculum-based tests. Hopfenbeck et al. (2016), for instance, demonstrate that PISA scores correlate strongly with overall intelligence, and Rindermann and Baumeister (2015) report that identified constructs are not consistent with assignments made by an expert group. Related research demonstrates moderate correlations with overall grades but low correlations between grades within subject areas (Pullkinen and Rautopuro, 2022; Fischbach et al., 2013). Both Liou and Bulut (2020) and Harlow and Jones (2004) demonstrate sensitivities in scores related to item format or testing method, with the former reporting on differences in scores when considering different item formats and the latter revealing differences in scores when comparing with interviews. Essentially, whilst it is understood that these tests certainly measure 'something', exactly what that something is and whether it is consistent with the stated construct are less certain.

It is generally accepted that for science, PISA, TIMSS, NAP-SL, and VALID are different tests, designed for different reasons, targeted at mostly different cohorts, and are therefore not directly comparable. However, comparisons between standardised tests are not uncommon, particularly when the comparisons are made for specific reasons (e.g. Wu, 2009). In this research, the focus is on assembling information about Australia's performance over time. Whilst achievement will be reported within the scope of each test, the compilation will provide a clearer picture of trends in student achievement in high school science.

## Methodology

### Data Sources and Analysis

This study involves comparing key indicators that have been already reported and undertaking new analyses on publicly available data sets. The existing reports this paper draws on include the report on Australia's 2018 PISA results (Thomson et al., 2019), the report on Australia's 2019 TIMSS results (Thomson et al., 2020), and the NAP-SL 2018 public report (ACARA, 2019). Any new analyses on is noted in the presentation of the results and was undertaken after download of publicly available Australian data available from ACER[1]. Data were analysed in SPSS in accordance with procedures set out in the corresponding technical reports (OECD, 2020).

[1] Available from https://www.acer.org/au/pisa/publications-and-data for PISA and https://www.acer.org/au/timss/reports-and-data for TIMSS

VALID data were accessed via a data request from the Centre for Educational Statistics and Evaluation (CESE), which is an arm of the Department of Education in the state of NSW. Additional ethics approval was not required (confirmed by UOW Human Research Ethics Committee). These data included over 45,000 individual student responses and approximately 500 NSW public schools. Data included just over 200 variables, including demographic data, individual item responses, scaled scores, and levels. Metadata including question coding (difficulty level, topic codes, syllabus codes) were also imported from a secondary data source. In this research, only student ID (unique, anonymous identifier), year, overall mark, overall percentage, overall scaled mark, and overall scaled band were utilised. Data from the years 2015, 2016, 2017, and 2018 were collected, cleaned, and combined into one dataset. Cleaning involved removing duplicates and inconsistencies in naming conventions of variables between years. Outliers were already removed from the data set prior to receipt. The data were assumed to be normally distributed. A Kolmogorov-Smirnov (K-S) test of normality was conducted on the overall percentage (OP) for each year (2015–2018) with all tests indicating skewed data ($p < 0.0005$). However, the K-S test for normality is sensitive to very large sample sizes and as such, the analysis of visual data, skewness, and kurtosis was carried out. Analysis from the Q-Q plots shows small deviation from normality for extreme scores. Further, skewness and kurtosis show values close to 0 (between − 1 and 1) which are indicators of approximate normal distribution. Homogeneity of variance was assumed due to equal group sizes.

## Key Indicators

In this paper, we are concerned only with student achievement as measured in standardised tests. In terms of student achievement, the key indicators include overall achievement scores and proficiency levels. Analyses related to these two indicators include overall achievement, performance over time (for overall achievement), proficiency (distribution of proficiency levels and proportion of students reaching minimum standards), and proficiency over time. These analyses are commonly explored with respect to comparisons between states, school sectors, sex, school location, indigenous status, and language background other than English status. Each test also reports or provides measures with respect to the key topic areas.

## Overall Achievement

Overall achievement is defined as a students' test score on the test as a whole. Overall achievement in PISA is measured by 'mean scores', which are often referred to as 'points' on a numerical scale called the scientific literacy scale. The scale has a mean of 500 for the international sample, with standard deviations of around 100 score points. One point different on this scale corresponds to an effect size of 1%. TIMSS follows the same definition though scores are simply referred to as mean scores or points. In NAP-SL, the average achievement score is similarly computed; however, for year 10 students, the mean score is 490 points. VALID scores are computed as both raw and scaled scores. The maximum scores for 2015 were 100 and from 2016 to 2018, the maximum score was 93. All raw scores are also reported as percentages. VALID assessments are calibrated and equated using the Rasch and partial credit models. The person ability in logits associated with each raw score for a given assessment is transformed into a scaled score with a mean of 85 and a standard deviation of 10. Comparisons of raw scores do not account for content and

difficulty differences between assessment items across assessment years. However, scaled scores are calculated after an assessment has been equated to the VALID scale, allowing comparisons across assessment years.

### Proficiency Levels

The tests reported in this paper approach proficiency levels in similar ways. PISA divides the scientific literacy scale into 7 proficiency levels, with 6 the highest and the lowest level 1, and is split into 1b and 1a, the former being the lowest. The scales are determined by a descriptive explanation of what students can do at each proficiency level, and they are then assigned a numerical value on the scientific literacy scale, depending on test analyses. Low performers are defined as student who scored below level 2 in scientific literacy (lower than 410 points), and high performers include students scoring at a level 5 or above. In Australia, the 'National Proficient Standard' is set at level 3, because this represents 'a reasonably challenging level of performance where students need to demonstrate more than the minimal skills expected' for 15-year-old students (ACARA, 2019). In TIMSS, four levels are identified: the advanced, high, intermediate, and low international benchmarks, set at 625, 550, 475, and 400 respectively. The intermediate international benchmark is also the National Proficient Standard, corresponding to a level 3 in PISA. In TIMSS, although the focus of this paper is on year 8 students, both year 4 and year 8 students are being assessed, and the proficiency levels apply separately to each (they are not the same scale such that year 4 students continue up the levels as they progress through school). For NAP-SL, however, there is a common proficiency level scale which applies to both cohorts sitting the test (year 6 and year 10). Altogether, there are five proficiency levels, with the proficient standard for year 6 set at the boundary between levels 2 and 3 and the proficient standard for year 10 set at the boundary between levels 3 and 4. NAP-SL proficiency levels are expressed as above or below the proficient standard for the year considered, rather than referred to as 'high' or 'low'. For VALID, there are six proficiency levels assigned. No studies have been undertaken to equate the VALID proficiency levels with national proficiency standards.

### Other Measures

Typically, in addition to overall achievement outcomes, reports for standardised tests will make comparisons between school region/location, gender, socio-economic status, indigenous status, language background, and immigrant status. International tests (TIMSS, PISA) are also able to make international comparisons. TIMSS, PISA, and NAP-SL also report on differences between states and PISA collects information about the school type (public, catholic, and independent). For each of the tests, a range of sub-scores can also be calculated, including those reflecting the different areas (e.g. knowledge or skills) or individual topics (e.g. biology, chemistry etc.). Results across the tests tend to be consistent for some measures (e.g. regions, socio-economic status, and indigenous status) and inconsistent for others (e.g. gender, language background, and individual constructs).

Currently, as Table 1 shows, there are no two tests which are identical with respect to population, design, content, or time period. For instance, NAP-SL and PISA overlap partially in terms of age (as some year 10s are 15 years old), but NAP-SL has only tested year 10s once, in 2018, so there is no data to draw from for comparisons over time. TIMSS and VALID both test year 8 students and both had administrations in 2019 (as VALID is a

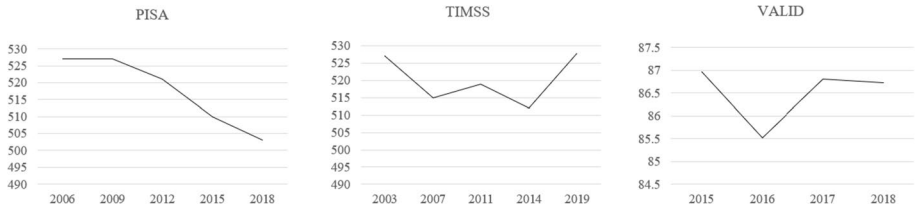


**Fig. 1** Trends in scores over time for PISA, TIMSS, and VALID. PISA scores for Australian 15 year old, TIMSS scores for Australian year 8 students, and VALID scores for all year 8 public school students in the state of NSW (note different time periods)

yearly test); however, VALID data includes only all public school students, whilst TIMSS does not collect information about school sector (e.g. public, independent).

## Results

### Overall Achievement and Cohort Achievement Over Time

Australian students scored an average of 502 score points in scientific literacy in PISA in 2018. As Thomson et al. (2019) note, this is above the OECD average of 489 score points. As previously mentioned, PISA shows a decline in mean scores over time (see Fig. 1). Statistically significant differences are reported between 2018 and all of the previous PISA cycles (2015, 2012, 2009, 2006). The report notes that 12 countries, including the UK and the USA improved their performance relative to Australia (e.g. USA and UK performed lower than Australia in 2006 but did not have statistically significantly different scores in 2018).

In TIMSS, Australian students scored an average of 528 points on the year 8 scale (Thomson et al., 2020). TIMSS scores have increased, declined, and then increased again over time, with statistically significant differences between 1995 and 2003, 1995 and 2019, 2003 and 2007, 2003 and 2015, 2007 and 2019, and 2015 and 2019 (Fig. 1).

The average score on NAP-SL in 2018 was 490 for year 10 students (± 7.3). Given this was the first administration to year 10, comparisons to previous years are not possible. However, the public report shows that there were no significant differences in year 6 scores between 2018 and 2015 or the cycles prior to this (ACARA, 2019).

VALID scores were computed, including raw scores (out of 100 in 2015 and 93 in 2016, 2017, and 2018), percentages, and scaled scores (Table 2).

Whilst raw scores reveal statistically significant differences between all years, scaled scores take into account cohort effects and will therefore be used in the year-on-year comparisons. A one-way ANOVA was conducted to determine whether differences in scaled scores over time were significant. The results indicate a significant effect [$F(3,182723) = 115.886$, $p < 0.001$]. Post hoc tests reveal differences between all years except 2017 and 2018, meaning that there was a statistically significant decrease in scores in between 2015 and 2016, followed by a statistically significant increase in scores between 2016 and 2017, as well as 2016 and 2018. The overall decrease across the period sampled, from 2015 to 2018, is also statistically significant.

**Table 2** Computed scores for VALID year 8 test

| | Year | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2015 | | | 2016 | | | 2017 | | | 2018 | | |
| | Raw | % | Scaled | Raw | % | Scaled | Raw | % | Scaled | Raw | % | Scaled |
| Total *N* | 45,299 | 45,299 | 45,059 | 46,059 | 46,059 | 45,738 | 46,026 | 46,026 | 46,026 | 47,243 | 47,243 | 45,904 |
| Mean | 51.77 | 51.77 | 86.97 | 52.22 | 56.15 | 85.52 | 50.68 | 54.49 | 86.80 | 48.79 | 52.46 | 86.73 |
| St. Dev. | 17.13 | 17.13 | 11.01 | 17.39 | 18.70 | 12.91 | 15.46 | 16.62 | 11.97 | 18.30 | 19.68 | 12.96 |

**Table 3** Proficiency level attainment computed for year 8 students for VALID 2015–2018

| | Individual levels | | | | | | Benchmarks | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | Level 4 and above | Low (1 and 2) | High (5 and 6) |
| 2015 | 0.91% | 8.19% | 30.55% | 45.14% | 14.55% | 0.66% | 60.35% | 9.11% | 15.21% |
| 2016 | 1.81% | 11.59% | 29.49% | 37.19% | 18.23% | 1.69% | 57.12% | 13.39% | 19.93% |
| 2017 | 0.20% | 10.40% | 28.58% | 39.62% | 19.78% | 1.42% | 60.82% | 10.60% | 21.20% |
| 2018 | 0.00% | 12.49% | 29.62% | 33.60% | 22.58% | 1.72% | 57.89% | 12.49% | 24.30% |

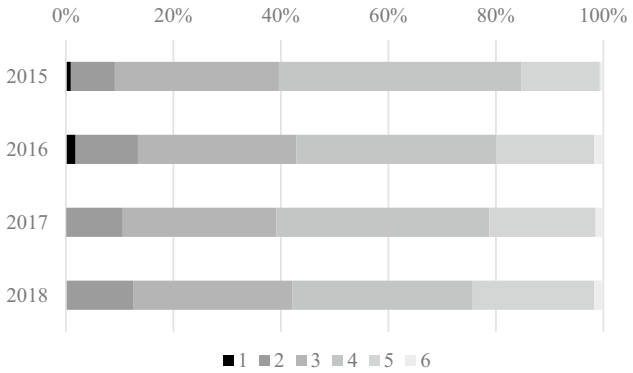


**Fig. 2** Proficiency level distribution for year 8 students for VALID 2015–2018

## Proficiency Levels and Cohort Proficiency Over Time

A second useful measure of achievement and achievement over time involves the examination of distributions across proficiency levels (see the 'Background' section).

PISA results show relatively stable distributions across proficiency levels between 2006 and 2012 and 2015 and 2018; however, there is an overall shift since 2006 with a larger number of students at lower proficiency levels and fewer at higher proficiency levels. That is, between 2006 and 2018, the proportion of high performers (levels 5 and 6) decreased by 5 percentage points and there was an increase in the proportion of low performers (levels 1 and 2) by 6 percentage points. Students achieving the National Proficient Standard (level 3) decreased from 2006 to 2018 by nine percentage points from 67 to 58%.

In TIMSS, the distributions across proficiency levels improve over time. There was a decrease from 1995 to 2019 (5 percentage points) of students in the low benchmark and an increase (3 percentage points) in the advanced benchmark. There were also differences between the years 2019 and all the other years from 1995. The proportions of students reaching the National Proficient Standard also increased by 5 percentage points from 69 to 74%.

The distributions in the VALID levels in 2018 are shown in Table 3 and Fig. 2. There are slight changes over time with more students (just over 3 percentage points) achieving a low benchmark (levels 1 and 2) and more students (an increase of 9

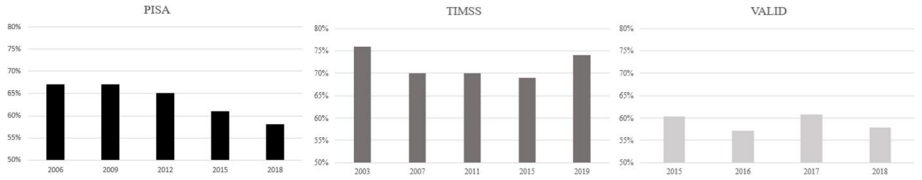


**Fig. 3** National Proficient Standard over time for PISA and TIMSS and level 4 and above attainment for VALID

percentage points) also achieving a high benchmark (levels 5 and 6). The proportion of students achieving at least a level 4 decreased from 2015 to 2018 by around 2.5 percentage point from 60.35 to 57.89%. Level 4 has not been established as the proficient standard equivalent in VALID but was considered to be the most appropriate, based on distribution alone. The decline was not linear, with decreases and increases between intervening years.

In NAP-SL, the proportion of year 10 students achieving the standard in 2018 (level 4 and above) was 49%. There are no data for NAP-SL to compare across years.

Figure 3 shows the trends in proficiency level achievement over time for PISA, TIMSS, and VALID.

## Additional Tests

Additional tests were also carried out to allow for comparisons which are closer in character. As VALID data include a whole sample of students in public schools, comparisons between PISA and VALID are possible by filtering out results by state and school sector (public only). The proportion of students attending government schools in 2018 NSW is around 65% (ABS, 2018)

In terms of PISA, comparisons could be made across the same time period (2015–2018). Note that the VALID test is administered to year 8 students, whilst PISA tests 15-year-old students (mostly in year 10). Strictly speaking, a closer comparison would involve comparing the same or overlapping cohorts, which require PISA data from intervening years (the VALID program commenced in 2014 but is based on the Essential Secondary Science Assessment (ESSA) which has been running from 2004 as a pen and paper test). The PISA dataset (2015–2018) was downloaded from ACER[2] and analysed in SPSS. Means were calculated on scaled scores, following technical notes (OECD, 2020). PISA mean scores for students enrolled in public schools in NSW in 2015 were 480.64 and 483.99 in 2018. There is no statistically significant difference between the 2015 and 2018 average scores [$t(3946) = -0.888$, $p = 0.374595$].

[2] https://research.acer.edu.au/ozpisa/43/

## Discussion

In this research, overall student achievement over time was explored with respect to four tests taken by Australian high school students, PISA, TIMSS, NAP-SL (Australia-wide), and VALID (NSW only). With respect to cohort performance over time, the results from existing and new analyses show an inconsistent picture. PISA reports a sharp decline in cohort scores over time (2008–2015) whilst TIMSS scores (2014–2019) show improvements over time. New analysis of VALID data (2015–2018) shows scores first improving and then falling over time, with a small but statistically significant decline over the period for which data was collected. The change is very small and potentially not significant when considering the sensitivity of statistical significance when working with large sample sizes (Khalilzadeh and Tasci, 2017).

The inconsistency between PISA and TIMSS results has been discussed in the wider literature (e.g. Wu, 2009). Explanations of the discrepancy focus on the fact that TIMSS is a curriculum-based test, where the items in TIMSS reflect expert group consensus on what students should know according to each nation's curriculum documents, whilst PISA is aimed at testing students' ability to *apply* their knowledge (not necessarily tied to any curriculum) to real-world contexts. Due to the different curricula experienced by students in different states and the significant differences in student achievement in the different states (e.g. ACARA, 2019), it is likely that either curricula or student demographics could be playing a role in over- or under-achievement. When PISA data were analysed with respect to NSW public schools only, to compare with VALID population groups, the declines present in the national sample are not evident. Tests comparing students with similar socio-economic advantage status in different states may be able to provide more insight into the nature of these differences, particularly before national policy and curriculum reform is enacted wholesale.

With respect to PISA, Australian students are performing worse not only in science but also in mathematics and reading. This supports empirical research showing intercorrelations between subject domains in PISA (e.g. Pullkinen and Rautopuro, 2022; Fischbach et al., 2013). These intercorrelations likely mean that PISA measures some general skill that is not associated strongly with the science literacy domain (e.g. Hopfenbeck et al., 2016). Whilst this might suggest that PISA science literacy scores are not necessarily instructive when measuring students' performance in science, they do still represent some other underlying construct. Studies which have analysed correlations between PISA scores (both separate by subject and overall) have shown that there are moderate correlations with a number of positive indicators, such as progressing through high school and high school grades (Pullkinen & Rautopuro, 2022). Overall, this indicates that whilst PISA scores might be important for measuring Australian students' overall proficiency, their utility for considering students' science knowledge is not well supported.

Turning to VALID, this test considers students' achievement related to outcomes as articulated by a single curriculum, the NESA Science Syllabus. The scores in VALID change slightly from year to year but with changes that are not remarkable. Whilst TIMSS is also a curriculum-based test, the content tested in TIMSS will not be as precise as VALID, as it is not as closely aligned to the assigned curriculum in that state. Australia does not have a single national curriculum, but rather, individual states either adopt the national curriculum (Australian Curriculum: Science) or 'adapt' the national curriculum to form state syllabuses, which is the case in NSW. Based on these observations, VALID scores over time do not seem to warrant concern. That is, students in NSW seem to be

achieving similarly in the period 2015–2018. Whilst it is feasible that curriculum reform might lead to improvements in students' scores, it is not accurate to say that students are achieving worse outcomes over time.

Proficiency levels across the different tests are arguably more valid with respect to comparisons, but, due to the nature of how the levels are defined, also more challenging. Trends in students' achievement of the National Proficient Standard vary across the tests but are consistent with trends in overall scores over time within each test. PISA, for instance, shows a decline in the number of students reaching the NPS whilst TIMSS shows a decline followed by an increase. Students reaching the VALID level 4 also decrease and increase over a 4-year period. Changes in the proficiency level attainment here seem to correspond to a change in the way these levels were measured; however, it was unknown at the time of publications what these changes were. For both TIMSS and VALID, changes in proficiency distribution involved a movement from the middle levels to both the lower levels and higher levels, whereas for PISA, shifts in level distribution tended to be unilateral, from higher to lower. No conclusions can be drawn therefore in relation to potential asymmetrical effects with respect to proficiency levels (i.e. it is not the case that either 'high' or 'low' tails were increasing or decreasing more so than any other group).

## Conclusion

Comparing scores across tests designed for different purposes has its limitations. The scores alone are not directly comparable, and even the most meticulously designed test will be inherently imperfect. However, these tests offer unprecedented insights into the all-important measure of student achievement and attempts to make meaning of the results of multiple tests is important, particularly for policy decisions. In this research, comparisons between tests demonstrate that concerns around decreasing achievement in science are not supported by the available evidence, including science-and-mathematics-specific test TIMSS and science-specific test VALID. Results from science-specific tests that are not consistent with declines reported in PISA support claims that PISA tests might be testing a more generic skill. Whilst this does not minimise the PISA result, it remains important for policy decisions, particularly around curriculum reform, which is often used as a blunt tool applied on a nationwide scale. More research is needed to identify whether curriculum effects can be separated from socio-demographic effects to ensure more targeted and effective approaches to educational reform.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions

## Declarations

**Conflict of Interest** The author declares no competing interests.

## References


ACARA, (2019). National assessment program science literacy 2018 public report https://nap.edu.au/nap-sample-assessments/results-and-reports

Araujo, L., Saltelli, A., & Schnepf, S. V. (2017). Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development, 19*(1), 20–34. https://doi.org/10.1108/IJCED-12-2016-0023

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. Routledge.

Education Council, (2015). *National STEM school education strategy: A comprehensive plan for science, technology, engineering and mathematics education in Australia*. https://www.education.gov.au/education-ministers-meeting/resources/national-stem-school-education-strategy

Fischbach, A., Keller, U., Preckel, F., & Brunner, M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences, 24*, 63–72. https://doi.org/10.1016/j.lindif.2012.10.012

Froese-Germain, B. (2010). The OECD, PISA and the impacts on educational policy. In *Canadian Teachers' Federation* (Vol. NJ1) http://files.eric.ed.gov/fulltext/ED532562.pdf

Hare, J. (2022). *Why Australia's students keep falling behind*. Australian Financial Review Online, Fairfax Media Management Pty Limited https://www.afr.com/work-and-careers/education/why-australia-s-students-keep-falling-behind-20220928-p5blna

Harlow, A., & Jones, A. (2004). Why students answer TIMSS science test items the way they do. *Research in Science Education, 34*, 221–238. https://doi.org/10.1023/B:RISE.0000033761.79449.56

Haugsbakk, G. (2013). From Sputnik to PISA shock–New technology and educational reform in Norway and Sweden. *Education Inquiry, 4*(4), 23222. https://doi.org/10.3402/edui.v4i4.23222

Hildebrand, J. (2023). *Our system is failing* (p. 7). Courier Mail.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Jerrim, J. (2023). Has Peak PISA passed? An investigation of interest in International Large-Scale Assessments across countries and over time. *European Educational Research Journal*. https://doi.org/10.1177/14749041231151793

Khalilzadeh, J., & Tasci, A. D. (2017). Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. *Tourism Management, 62*, 89–96. https://doi.org/10.1016/j.tourman.2017.03.026

Kane, M. T. (2016). Explicating validity. *Assessment in education: Principles, policy & practice., 23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Liou, P. Y., & Bulut, O. (2020). The effects of item format and cognitive domain on students' science performance in TIMSS 2011. *Research in science education, 50*(1), 99–121.

NSW Education Standards Authority (NESA). (2020). *Nurturing wonder and igniting passion, designs for a new school curriculum: NSW Curriculum Review* https://nswcurriculumreform.nesa.nsw.edu.au/home/siteAreaContent/524abec1-f0f9-4ffd-9e01-2cc89432ad52

OECD. (2020). *PISA 2018 technical report*. OECD Publishing https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD, (2019). *PISA 2018 results (volume 1): What students know and can do*. https://www.oecd.org/education/pisa-2018-results-volume-i-5f07c754-en.htm

Office of the Chief Scientist. (2014). *Science, technology, engineering and mathematics: Australia's future*. Australian Government ACT: https://www.chiefscientist.gov.au/2014/09/professor-chubb-releases-science-technology-engineering-and-mathematics-australias-future

Pons, X. (2017). Fifteen years of research on PISA effects on education governance: A critical review. *European Journal of Education, 52*(2), 131–144. https://doi.org/10.1111/ejed.12213

PricewaterhouseCoopers (PwC) (2015). *A smart move*. https://www.pwc.com.au/pdf/a-smart-move-pwc-stem-report-april-2015.pdf

Pulkkinen, J., & Rautopuro, J. (2022). The correspondence between PISA performance and school achievement in Finland. *International Journal of Educational Research, 114*. https://doi.org/10.1016/j.ijer.2022.102000

Rindermann, H., & Baumeister, A. E. (2015). Parents' SES vs. parental educational behavior and children's development: A reanalysis of the Hart and Risley study. *Learning and individual differences, 37*, 133–138. https://doi.org/10.1016/j.lindif.2014.12.005

Thomson, S., De Bortoli, L., Underwood, C., & Schmid, M. (2019). PISA 2018: Reporting Australia's results. Volume I student performance. In *Australian Council for Educational Research (ACER)* https://research.acer.edu.au/ozpisa/35

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

Waldow, F. (2009). What PISA did and did not do: Germany after the 'PISA-shock'. *European Educational Research Journal, 8*(3), 476–483. https://doi.org/10.2304/eerj.2009.8.3.476

Wiseman, A. W. (2013). Policy responses to PISA in comparative perspective. In H. D. Meyer & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 303–322). Symposium Books.

Wu, M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Prospects, 39*, 33–46. https://doi.org/10.1007/s11125-009-9109-y

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change, 21*(2), 245–266. https://doi.org/10.1007/s10833-019-09367-x

## Further Reading

Australian Bureau of Statistics. (2022). *Schools ABS* https://www.abs.gov.au/statistics/people/education/schools/latest-release.

Global business coalition for education and the education commission (2019): *The 2030 skills scorecard: Bridging business, education, and the future of work*. https://gbc-education.org/wp-content/uploads/sites/2/2019/09/GBC-Education-2030-Skills-Scorecard.pdf