Australian Education Research Organisation

Benchmarking performance

Future directions for Australia's National Assessment Program

March 2023

The Australian Education Research Organisation is Australia's national education evidence body, working towards excellent and equitable outcomes for all children and young people.

Acknowledgements

The Australian Education Research Organisation (AERO) acknowledges the traditional custodians of the lands, waterways, skies, islands and sea country across Australia. We pay our deepest respects to First Nations cultures and Elders past and present. We endeavour to continually value and learn from First Nations knowledges and educational practices.

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

Authors: Daniel Carr, Laura Good, Lihini De Silva, Jenny Donovan, Zid Mancenido

Other contributors: Kate Ridgway, John Ainley

Cover image: monkeybusinessimages (iStock)

How to cite

Australian Education Research Organisation (2023), Benchmarking performance, edresearch.edu.au

Further information

AERO produces resources to support the use of high-quality research. Explore these at edresearch.edu.au/browse.

Contents

Glossary of terms and acronyms 4

Introduction 5

1. Rationale for a National Assessment Program (NAP) 6

- Introduction 6
- Components of the NAP 6
- The national context 7
- Does the NAP achieve its purpose? 8
- Expectations of NAPLAN 8
- Summary 10

2. What national assessment results tell us 11

- Divergent trends in the NAP assessments 11
- Can divergent NAP results be explained? 12
- Summary 16

3. Making use of assessment data to understand ‘what works’ 17

- What assessment data can tell us 17
- The limitations of using assessment data to identify ‘what works’ 19
- Summary 23

References 24

Appendix A: Further detail on the National Assessment Program 28

Appendix B: Preliminary analysis conducted by the ACER 32

Glossary of terms and acronyms

ACARA	Australian Curriculum, Assessment and Reporting Authority
ACER	Australian Council for Educational Research
IEA	International Association for the Evaluation of Educational Achievement
KPM	Key performance measures
LSAY	Longitudinal Surveys of Australian Youth
Framework	Measurement Framework for Schooling in Australia
NAP	National Assessment Program
NAPLAN	National Assessment Program – Literacy and Numeracy
National Report	National Report on Schooling in Australia
NPD	National Pupil Database
NSRA	National School Reform Agreement
OECD	Organisation for Economic Co-operation and Development
OFAI	Online Formative Assessment Initiative
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
Standardised assessments	Tests that are administered and scored in a predetermined and consistent way
SES	Socioeconomic status
TIMSS	Trends in International Mathematics and Science Study

Introduction

Over the past three decades Australia has developed an increasingly advanced national system of student assessments, results from which have been used to identify areas of growth, stagnation or decline in student learning. For the most part, trends in different standardised assessments have been considered in isolation. By examining literacy and numeracy results across assessments, we can better understand the performance of Australian students over time; we can pinpoint areas of national strength and weakness and improve Australia's educational outcomes.

This report considers the four National Assessment Program (NAP) assessments that measure literacy and numeracy:[1] the National Assessment Program – Literacy and Numeracy (NAPLAN), Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA). NAPLAN is conducted by the Australian Curriculum, Assessment and Reporting Authority (ACARA) and assesses how students are progressing over time, while monitoring system-level and school-level performance. The other three assessments – PIRLS, TIMSS and PISA – are international programs that all jurisdictions chooses to participate in, to benchmark the learning outcomes of Australian students against their peers in countries around the world.

First, this report examines the purpose of the National Assessment Program. It finds that, over time, many purposes have been ascribed to the NAP and NAPLAN in particular, and suggests that this may have created some confusion and undermined confidence as to whether the assessments are fit for purpose.

Second, this report examines reasons why PISA shows significant declines in both reading and mathematics achievement, while NAPLAN, PIRLS, and TIMSS show either growth or stability. Drawing on preliminary analysis by the Australian Council for Educational Research (ACER) on behalf of AERO, this report finds no single cause can be definitively identified.

Finally, the report explores how the NAP assessments have the capacity to tell us much about effective practice and policy; they can help to detect 'what works' in education. While there are some limitations, analysis of NAPLAN, PISA, PIRLS and TIMSS trends can help to identify policies and practices that may have contributed to improvements over time. In addition, limitations could be addressed, in part through data linkages and the creation of a central student data set, and by surveying students and teachers when NAPLAN is conducted to provide richer detail on the classroom practices and school approaches being used.

The National Assessment Program is an important investment made by all Australian governments. It is an asset that helps measure the health of Australian school education. This report considers how the usefulness of the NAP can be enhanced to improve our evidence base about the successes and challenges of the Australian school system. It is timely to do so given Australia's suite of national assessments have been in place for more than a decade and the Measurement Framework for Schooling in Australia (Framework), which is used to measure school system performance, is currently under review.

1 The broader NAP includes the yearly NAPLAN assessments, the 3-yearly sample assessments in science literacy, civics and citizenship, and information and communication technology (ICT) literacy, and the international sample assessments PIRLS, TIMSS and PISA.

1. Rationale for a National Assessment Program (NAP)

Introduction

Australia's system of national assessments, the National Assessment Program, has been in place for more than two decades.[2] Elements of it have changed over time – most notably the addition of the standardised national assessments of literacy and numeracy known as NAPLAN. The NAP has been the basis for Australia's education performance monitoring and benchmarking. It has often attracted criticism: sometimes motivated by a general opposition to standardised testing, or concern about technical aspects of test design and administration; sometimes arising from concern that the assessments are not fit for purpose or do not adequately meet the needs they are intended to address.

This report provides an overview of Australia's approach to using standardised assessments to measure and benchmark school system performance, and to understand 'what works' in schooling. It primarily concentrates on literacy and numeracy, reflecting their foundational status in schooling.[3]

The report is structured in 3 sections.

Section 1 explores the rationale for the design of the NAP and considers whether the current NAP assessment mix is able to meet emerging demands from policymakers.

Section 2 comments on the NAP results and identifies factors that might be driving the divergence in performance across various measures.

Section 3 details how data from the NAP assessments offer insights into 'what works' in teaching, policy and programs, before identifying the limitations to use of the assessments for this purpose and suggesting some solutions.

Components of the NAP

Australia uses four main standardised assessments to measure literacy and numeracy achievement over time, which in part comprise the National Assessment Program (NAP).[4]

Literacy and numeracy components of the National Assessment Program

The National Assessment Program – Literacy and Numeracy (NAPLAN) is an Australian assessment of literacy and numeracy skills aligned to the national curriculum in Years 3, 5, 7 and 9. This assessment tracks how a learner is progressing against national standards over time. All students are expected to sit the tests but may be exempted or withdrawn for certain reasons. It has been administered every year since 2008, except for 2020.

The Programme for International Student Assessment (PISA) is an international assessment of 15-year-olds' ability to apply their knowledge and skills to real-life problems and situations, focusing on reading, mathematics and science. A nationally representative sample of more than 14,000 Australian students in over 700 schools complete the test. It has been administered every 3 years since 2000, with the 2021 test delayed until 2022.

The Trends in International Mathematics and Science Study (TIMSS) is an international study of mathematics and science achievement in Years 4 and 8. A representative sample of students from each Australian jurisdiction, school sector, location and socioeconomic status sit the test, with about 6,000 students participating in Year 4 and 9,000 in Year 8. It has been administered every 4 years since 1995, with the next test due to be conducted in 2023.

2 Before NAPLAN, there existed a range of standardised assessments of literacy and numeracy conducted at jurisdictional level, and some sample-based standardised assessments conducted periodically at the national level.

3 In this report, literacy is measured with reference to reading given it is the one common domain across the nationally collected assessments. The focus of numeracy varies across these assessments, as do the terms used to describe this domain. To avoid confusion, alternative terms (for example, 'mathematical literacy' and 'mathematics') are not used, except for when the specifics of what is tested in each assessment are discussed in Section 2.

4 Additional NAP sample assessments for science literacy, information and communication technology, and civics and citizenship also occur domestically on a rolling 3-year basis. While not analysed in this report, further investigation of these sample assessments could provide further insights into the affordances of sample-based assessments as well as issues related to common-item equating across year levels and over time.

The Progress in International Reading Literacy Study (PIRLS) is an international study of Year 4 reading literacy achievement, with about 6,000 Australian students participating across 280 schools. It is administered every 5 years, with Australia first participating in 2011 and the latest test in 2021.

Further information about each of these assessments is available in Appendix A.

The national context

From the Hobart Declaration in 1989 to 2019's Mparntwe (Alice Springs) Declaration, there have been regular efforts to describe agreed education objectives between national, state and territory governments (OECD 2011:124). The creation of national institutions and a series of intergovernmental agreements setting out shared aspirations, targets and policy reform commitments have been intended to deliver the shared outcomes.

Following years of state-specific standardised assessment collections, the National Assessment Program was established as an outcome of the 1999 Adelaide Declaration. The NAP was intended to gather, analyse and communicate student achievement data in a nationally comparable and transparent way (ACARA 2016a).

Two decades on, the NAP remains the main mechanism for monitoring student achievement in key learning domains. The Mparntwe Declaration (2019:5–9) built on themes of school system monitoring, accountability and improvement that had been described in previous declarations. It established the overarching goal of promoting excellence and equity in education, and outlined further ambitions, such as 'promoting world-class curriculum and assessment' and ensuring 'Australia's education system is recognised internationally for delivering high quality learning outcomes.' The Mparntwe Declaration also noted the importance of 'good quality data' to measure and benchmark system performance, and to collect evidence on 'what works'. In addition to these 'assessment of learning' purposes, there is also a commitment to developing and enhancing assessment as and for learning (effectively, to facilitate student self-reflection and to inform teaching practice), and to providing data on student and school outcomes to parents and carers for accountability purposes.

The National School Reform Agreement (NSRA), which was developed jointly by the (then) Council of Australian Governments (2018), establishes the objective of achieving high-quality and equitable education, and related outcomes such as improving achievement for all students. It also sets out shared policy initiatives, including enhancing the national evidence base. Like the Mparntwe Declaration, the NSRA also commits to creating a school system that is recognised internationally for its learning outcomes.

The goals articulated in the Mparntwe Declaration and NSRA inform the Measurement Framework for Schooling in Australia (ACARA 2020a), which sets out nationally agreed key performance measures (KPMs) and their corresponding data sources. Progress towards these KPMs is reported annually in the National Report on Schooling in Australia. The various literacy and numeracy assessments that make up the NAP form the basis of a number of KPMs in the Framework.

The Education Council of ministers has described the NAP as important in enabling monitoring of achievement across the country, benchmarking against other nations and identifying effective teaching practices; all of which align with the Mparntwe Declaration and NSRA (ACARA 2016a; MCEETYA 2009).[5]

The NAP, including its NAPLAN component, is also described by ACARA (2016a, 2016b) as meeting other purposes beyond those listed above. These include providing:

- data for school and system accountability
- an input (and outcome) for school improvement planning
- information about student attainment to inform teaching
- parents, carers and teachers with information about student and school performance.

5 NAP results can also be used for other purposes, such as in Victoria where additional funding is provided to schools on the basis of the number of students falling below the national minimum standard in Year 5 NAPLAN reading (Department of Education and Training Victoria 2021).

Does the NAP achieve its purpose?

There are three main considerations here.

The first is that the suite of assessments in the NAP can be perceived as achieving its purpose to *some extent*. The suite of assessments in the NAP does provide insights into aspects of Australia's education system, enabling monitoring and benchmarking of learning achievement, within limits. Each of the assessments differ in what is measured, the age groups tested, the point at which students are tested and the frequency of the testing. The results from each are reported separately as each test has its own scale, making it difficult to directly compare findings across assessments. As assessments of student learning, they can give a disjointed and sometimes seemingly contradictory story about the learning achievement of students (this is explored further in Section 2).[6]

The second consideration is that, at the same time as the assessments may be seen as limited in delivering their main purpose – to measure and benchmark student learning progress over time – the assessments are also underestimated as important sources of insight into the performance of our education policies and practices. The suite of assessments provides more than just test responses. It also delivers information from students, teachers and others that could enable better monitoring and evaluation of aspects of our education system, but we do not collect or use this data systematically. This unfulfilled potential to meet the broader purpose of the NAP is explored more in Section 3 of the report.

Third, there is evidence that the 'purpose' of the NAP has become contested. There are apparent gaps between stated purpose and assumed purpose, meaning that stakeholders (ministers) may believe that the existing program can give insights it is not designed to provide. This confusion of purpose is particularly evident in relation to NAPLAN.

Expectations of NAPLAN

NAPLAN is the most informative of the assessments, as it is conducted annually, designed to assess full cohorts of students in Years 3, 5, 7 and 9, and it examines aspects of both literacy and numeracy. However, NAPLAN's purpose has changed over time, and it has been vulnerable to attempts to overextend the insights it can provide.

The McGaw et al. (2020) review of NAPLAN noted five distinct purposes currently being claimed for NAPLAN. These include:

1. monitoring – examining progress towards national goals for education and tracking policy impact
2. accountability – transparent reporting of results at the school, sector and jurisdiction level
3. school improvement – using the data to identify strengths and weaknesses at the school level and target programs and interventions to help schools lift their results
4. individual achievement – putting individual results in national context and providing data on students' skills that can be triangulated with teacher judgement and other assessments
5. information for parents/carers – providing individual and school-level data on achievement and growth in terms of bands of achievement.

These largely agree with the purposes of NAPLAN stated by ACARA. However, McGaw et al. (2020) acknowledged there is a lack of consensus among stakeholders as to these purposes and whether they can reasonably be achieved within a single standardised assessment like NAPLAN. In particular, concern was raised that NAPLAN 'was designed for system level data' but is being used 'for individual student data' and/or to compare and judge specific schools (McGaw et al. 2020:26–27).

It may not be possible for one assessment to effectively deliver five different functions.[7] The debate as to whether and how NAPLAN serves as a diagnostic assessment tool is illustrative of the need for a clear purpose to inform design.

6 We do not wish to imply that the goal is for NAP assessments to tell a uniform story; rather, apparent contradictions can be informative for policymaking if investigated further, as we explore in Section 2.

7 There has been less confusion about the purpose of the other NAP literacy and numeracy assessments (PISA, TIMSS and PIRLS), as their sample-based collection limits demands for their wider use.

Some have long ascribed a diagnostic purpose to NAPLAN. Then education minister Julia Gillard (Commonwealth of Australia 2010:22) said about NAPLAN:

> It is important to teachers; they do value this diagnostic information to work out what they need to do next for the children in their class.

While former ACARA chair, Professor Barry McGaw (Commonwealth of Australia 2014:41), claimed:

> NAPLAN is not a test students can prepare for because it is not a test of content. The federal government's intention in introducing and reporting NAPLAN results was to provide a diagnostic tool for teachers and parents, identifying gaps in students' skills.

However the then CEO of ACARA, Dr Peter Hill (Commonwealth of Australia 2010:22), noted the limits to NAPLAN being truly diagnostic:

> Diagnostic assessment means that we look at the reasons why students are, perhaps, not performing. For that purpose we need immediate feedback; these tests are broad in scope and would not be very useful for diagnostic purposes, particularly as the results come through very late.

This view was echoed more recently by NSW Education Minister Sarah Mitchell (Baker and Cook 2019):

> In 2019, it is clear that a diagnostic test must be on demand, it must be linked to the curriculum, it must focus on student growth, and it must test informative writing. NAPLAN in its current form does not meet [these] criteria.

As highlighted in the Senate inquiries in 2010 and 2013–14, when assessed objectively, NAPLAN is most suited to the purposes of supporting system-wide policy decisions, school improvement, identifying trends by comparing results each year and enabling parents and carers to track student performance. As one submission to the NAPLAN review stated, 'It is difficult to simultaneously achieve census and system testing in conjunction with diagnostic testing for teachers' (McGaw et al. 2020:27).

Though efforts have been made to support teachers to use NAPLAN data diagnostically, they have seen limited success. One such initiative was the Victorian Curriculum and Assessment Authority's (2013) development of resources to evaluate student performance using NAPLAN and to plan their teaching and learning programs using the results. Another more recent example is the development of the insights packages from the NAPLAN writing data that identify strengths and weaknesses in student writing to help inform teaching decisions in schools in NSW (CESE 2019). However, these resources were unable to resolve the main problems that teachers have with using NAPLAN diagnostically, namely the difficulty of pinpointing particular areas of student weakness given the span of the test and the time taken to release the results.

Further, the delay between when students have previously sat the test (in May) and when results are released (generally August to September) limited the value that teachers place on using NAPLAN to inform their teaching (Kostogriz and Doecke 2011; Rogers et al. 2018). Instead, many teachers perceive NAPLAN's purpose as providing accountability and benchmarking (Polesel et al. 2014).

ACARA has previously endeavoured to eliminate the confusion around whether NAPLAN was intended as a form of assessment for learning by noting that an assessment can provide diagnostic value at the school rather than student level. At the 2010 Senate inquiry, ACARA (2010) clarified:

> NAPLAN is not a diagnostic assessment for the individual student … However, there is another sense where the use of the term diagnostic assists a general audience to understand the principle of useful data to evaluate teaching and learning programs … In this sense therefore, NAPLAN is 'diagnosing' the strengths and weaknesses of schools' teaching and learning programs and informing future programs, by identifying gaps in student knowledge and skills.

Upcoming changes to NAPLAN may provide an opportunity for teachers to better use NAPLAN data in a formative or diagnostic way. The education ministers have agreed that, from 2023, NAPLAN should be conducted earlier in the school year (in Term 1). This change, and the recent move to universal online delivery, means that reports on student performance can be provided earlier, and used as formative assessment[8] by teachers (Education Ministers 2022).

Whether or not NAPLAN was ever intended to be diagnostic, it is clear that teachers and systems have not viewed NAPLAN as sufficiently diagnostic for their purposes. This is why many jurisdictions have developed additional system-wide assessments that are designed for formative purposes. For example, the reading and numeracy 'Check-in' assessments provided in NSW for students in Years 3 to 9 have been mapped to the National Literacy and Numeracy Learning Progressions, with results delivered shortly after the completion of the assessment. This enables teachers to identify student performance and tailor their planning to student needs, with additional resources on teaching strategies also provided in the portal where they receive student assessment feedback (NSW Department of Education 2022).

This diagnostic purpose was also implicit in the rationale for the Online Formative Assessment Initiative (OFAI), a national initiative in the current NSRA. The OFAI was intended to support teachers in using formative assessment. It was designed to give teachers a way of collecting and recording assessment data, as well as providing a suite of assessment tools and professional learning resources on formative assessment (OFAI 2020). At their meeting in December 2022, education ministers decided to halt further development of the OFAI and instead agreed to adapt existing NSW and Victorian formative assessment resources so they are available to all teachers.

Summary

In the NAP, Australia has a suite of assessments that currently only meets the intended purpose of monitoring and benchmarking student learning achievement to a limited extent. At the same time, it is clear that stakeholders (ministers) have an appetite for diagnostic assessments that will support teachers to use information about student learning formatively. The NAP assessments are not designed to meet this purpose.

8 That is, teachers can use the data to inform their planning and teaching, targeting areas that need more attention.

2. What national assessment results tell us

To improve Australia's educational outcomes, we need to understand student performance over time, to identify areas of growth, stagnation or decline, so that we can prioritise attention and resources. The NAP assessments are the primary means of understanding achievement at a system level, yet they tell different stories about the performance of students over time. AERO found that NAPLAN, PIRLS and TIMSS show either growth or stagnation, while PISA shows significant declines in both literacy and numeracy achievement. This injects a degree of uncertainty into the picture of system-level literacy and numeracy achievement and progress over time.

This section explores the disparate trends across the NAP assessments before considering whether the apparent divergence is unique to Australia. A preliminary investigation by the Australian Council for Educational Research (ACER), on behalf of AERO, considered possible explanations for the different trends in NAP assessments (see Appendix B for further information). This includes statistical and sampling issues and differences in the assessments relating to factors such as their format, design, style of questions asked, content and curriculum coverage. More research is needed to form conclusions about what is causing the divergence between PISA and the other NAP assessments.

Divergent trends in the NAP assessments

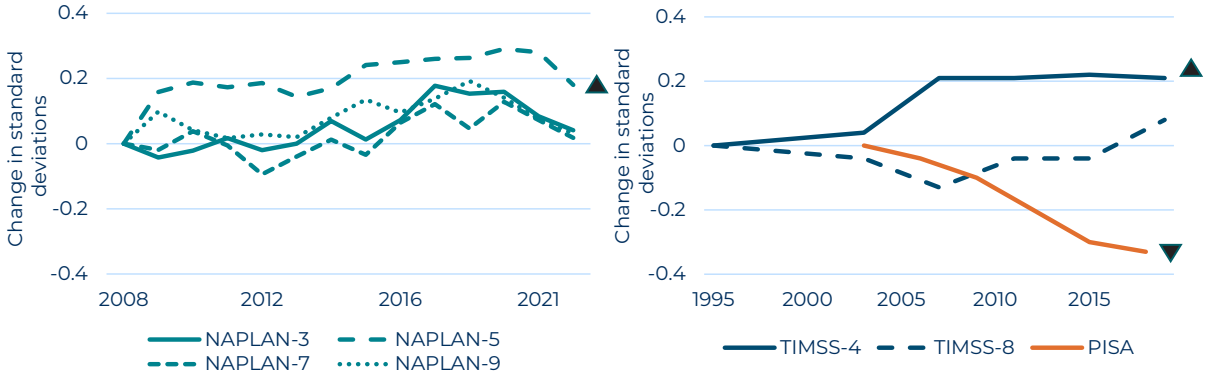
NAP assessments do not tell a consistent story about student achievement in either literacy or numeracy. PISA is the outlier. NAPLAN, PIRLS and TIMSS results show either upward trends or stasis. On the other hand, PISA results show a significant decline since the test was first administered in the early 2000s.

Figure 1 and Figure 2 provides a visual account of these trends.[9] To compare trends across these assessments, AERO calculated the change in average achievement for each assessment between a given year and the baseline year. This change is measured in standard deviation units (a measure of variation) and is reflected in the vertical axis of Figure 1 and Figure 2.

The stark differences in trends paint a contradictory picture of students' performance over time. Despite this, there has not been a significant effort to understand what might explain these mixed results, or an assessment of whether any measure is a more reliable and informative indicator of future educational and employment outcomes.

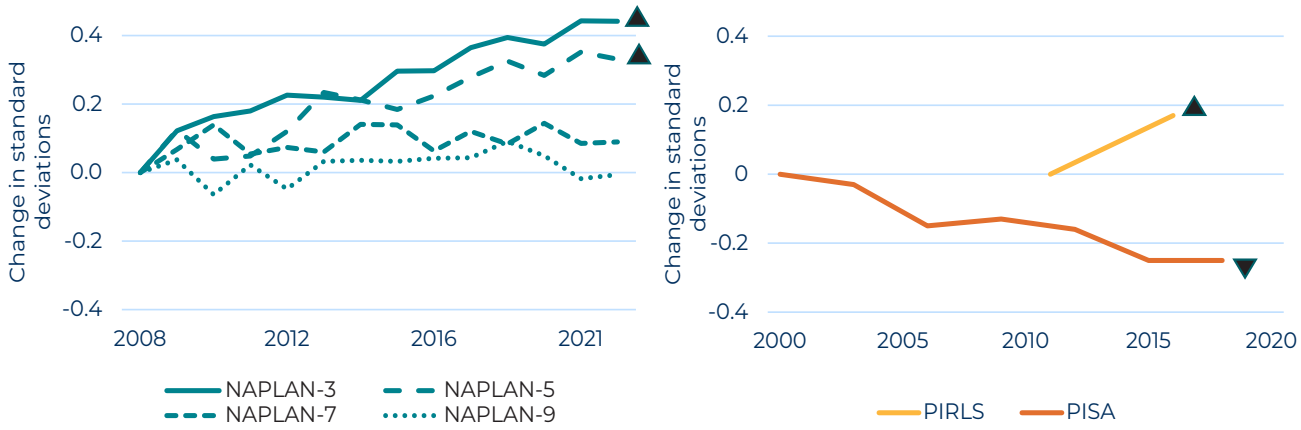
There is also a significant gap between students from low and high socioeconomic status (SES) backgrounds. According to AERO's analysis of NAPLAN data, this gap has increased over time.

Figure 1: Standardised achievement in numeracy over time



9 NAPLAN literacy achievement in Figure 2 is calculated using reading scores only.

Figure 2: Standardised achievement in literacy over time



Note: "Change in standard deviations" refers to the change in the average score of Australian students on the test relative to the base year (e.g., NAPLAN was first administered in 2008). Change was calculated by transforming the average scaled scores of each test into standard deviation units relative to the first administration of the test. An upward pointing symbol (▲) denotes a statistically significant increase in the average score in the latest year the test was conducted (up to 2022) relative to the first year. A downward pointing symbol (▼) denotes a statistically significant decrease. If there is no symbol, there was no statistically significant change.

Source: AERO analysis of data supplied by ACARA and ACER.

For example, achievement in numeracy has increased significantly for students who have a parent with a bachelor's degree, while it has decreased for students whose parents have a Year 11 education or less.[10] PISA shows some decrease in this gap, with a slightly greater decline in achievement from the highest SES quartile than the lowest over time. It should be noted that socioeconomic status is measured differently in NAPLAN and PISA, with NAPLAN reporting achievement by parental education based on parent self-reports, while PISA reports achievement in terms of SES quartiles on an economic, social and cultural status index (ESCS) based on student self-reports.[11]

Can divergent NAP results be explained?

There are key distinctions between the NAP assessments that may be important for interpreting trends and gaining an insight into why PISA results have diverged from the other assessments.

A preliminary investigation conducted by ACER on behalf of AERO identified the following as potentially relevant factors[12] worth exploring:

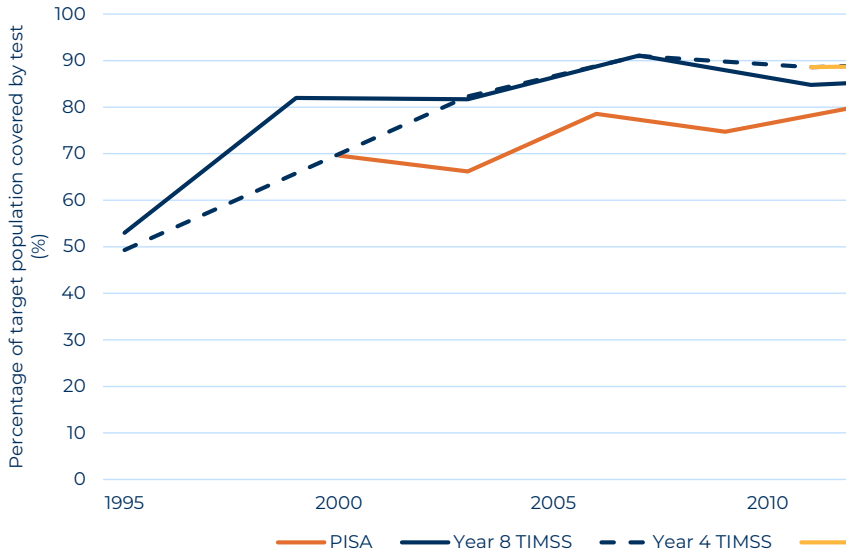
- assessment participation (including sampling methods and student exemption policies)
- scaling and equating processes
- assessment format (particularly whether the assessment is conducted online or when it was moved online)
- assessment design (including whether the assessment is adaptive)
- question design (including reading load and degree of abstraction)
- content and curriculum coverage (including what constructs are being measured).

10 This analysis is based on publicly available data. It is important to note that as parental education information in NAPLAN is self-reported, further analysis is needed into how student participation and parental response rates have changed over time, and how this may impact interpretations of trends.

11 It may be possible to better understand the relationship between SES and achievement over time by investigating whether there are similar items used across NAP assessments to measure SES.

12 There are other factors that could be investigated in future, such as number of items in each assessment and the relative difficulties of items across assessments. However, these factors were prioritised given they are the most commonly referenced potential explanations for divergent results.

Figure 3: Trends in potential sampling bias across international assessments



Note: The percentage of target population covered by test is calculated by considering the proportion of the target cohort enrolled in the school system, the student-level exclusion rate, the extent to which schools refuse to participate, and student absences. The percentage is computed as follows: 100%−[[100%−(proportion of population enrolled in school)] + (proportion of students excluded) + [100%−weighted (school response rate after replacement)×(student participation rate)]]. Anders et al. (2021) discusses how this measure of participation is holistic and enables comparisons across assessments (which may have different exclusion policies).

Source: AERO analysis of data from OECD, IEA and ACER.

Assessment participation

There are differences between the NAP assessments in who participates in them due to their different designs. NAPLAN assesses all students in target year levels to collect data for reporting on individual students and schools (a census approach). In contrast, the international assessments focus on education systems, so they are administered to a sample of target students. There are differences among the NAP assessments in their student and school sampling practices,[13] and NAPLAN applies student exemption policies.[14] However, these settings have remained fairly constant over time, so they do not account for the divergence in test outcomes.

Most of the NAP assessments have seen some change in participation rates over time. Participation rates are important as literature suggests they can influence system-level results (Anders et al. 2021; Jerrim 2013, 2011; Pereira 2011).

Pereira (2011) argued that changes to how Portugal's PISA sample was drawn influenced results between 2003 and 2009, with shifts in the distribution of students by year level being a key contributor. Figure 3 shows the extent to which the international assessments have managed to cover the target populations in Australia over time. The higher the proportion of the target population covered by the assessment, the lower the bias that may artificially inflate reported test results. Figure 3 shows that coverage has increased over time across PISA, TIMSS and PIRLS. This suggests that changes in participation rates cannot fully explain divergence in PISA results.

13 For example, one difference between TIMSS/PIRLS and PISA in sampling students within sampled schools is that TIMSS/PIRLS randomly sample intact class(es) within each school whereas PISA randomly samples individual students from a list of all eligible students within the school. This means that design effects (clustering) are smaller for PISA than TIMSS.

14 While a range of adjustments to the NAPLAN tests are available for students with disability, some students with disability may be exempted from sitting the tests. Similarly, students with a language background other than English and limited English language proficiency may also be exempted from the tests. Details on the exemption policies are available at nap.edu.au/naplan/for-schools/student-participation.

For NAPLAN, participation rates have fallen over time, from about 97% in 2008 to 92% in 2022 across Years 3, 5 and 7.[15] Year 9 rates started lower at 93% and have declined to about 87% in 2022 (ACARA 2022). In that year, more than 37,000 Year 9 students and close to 24,000 Year 7 students did not participate in the test (recorded as absent and withdrawn), with about 70% of non-participation due to students being absent on the day of the test.[16] Similar to the participation pattern observed for international assessments, students who did not participate in NAPLAN tended to be lower performers (see, for example, CESE 2016). NAPLAN mean scores reported at a national and subnational level are adjusted to take into account missing data resulting from non-participation, by a process known as 'imputation'. This process uses data of like students (for example, similar socio-educational background and enrolled in similar schools) to predict the scores of those who were absent or withdrew from the test. However, this process may overestimate the performance of missing students, thereby inflating the reported means.[17]

Ainley et al. (2020) explored the possibility that shifts in the age-grade distributions of students in the Australian PISA sample may have contributed to the decline in scores seen over time. Australian data from all PISA cycles show that Year 11 students receive higher scores than their Year 10 peers, who in turn score more highly than Year 9 students.[18] In the 2018 PISA cycle, 12% of Australia's participants were in Year 9 (up 6 percentage points since 2000), 81% were in Year 10 (up 4 percentage points since 2000) and 7% were in Year 11 (down 10 percentage points). Overall, the analysis suggests that these shifts cannot wholly explain Australia's declining PISA achievement, as changes in achievement have not always corresponded with shifts in the year-level distributions of students (Ainley et al. 2020).

This analysis also shows that the PISA literacy and numeracy scores of Year 9 students did not change significantly between 2000 and 2018 (that is, the overall drop is a product of declining results among the Year 10 and Year 11 students, alongside the shift in year-level distributions) (Ainley et al. 2020). This accords with the results of Year 9 NAPLAN tests, which also show no significant change between 2009 and 2022 (see Figure 1 and Figure 2).

Overall, trends observed about which students sit the assessment do not tell a definitive story that can explain why PISA has diverged from the other NAP assessments. Further research may be able to shed more light on this matter, as it has in other countries such as Portugal and the United Kingdom. The one remarkable finding is that the performance of Year 9 students in PISA has not declined, which suggests that changes in schooling that have most affected Years 10 and 11 may be worth exploring further.

Scaling and equating processes

Each NAP assessment uses scaling and equating models. Scaling is used for measurement accuracy. and to enable longitudinal comparisons. Equating is done to adjust the results of each test so they are comparable to previous years' data, as tests may be easier or more difficult than other years.

Although all four assessments use the same metrics – a mean of 500 and standard deviation of 100 – they are not comparable, due to the different selection of countries they include in their sample and the distinctions in their conceptualisation, operationalisation and content coverage.

15 For the purpose of calculating participation rates, participating students include exempt students but not those who were absent or withdrawn by their parents. Rates quoted in this section are those averaged across all tests.

16 NAPLAN participation rates vary significantly between states and territories, which may also complicate between-jurisdiction comparisons over time. For example, 1 in 4 students in the NT and 1 in 6 in Qld did not participate in the Year 9 reading test, much higher than the 5% non-participation rate in WA and 6% in NSW. Qld and NT also had the greatest average annual decline rate of all jurisdictions over the past 5 years. The high Year 9 participation rate in WA may partly be due to the use of the Year 9 NAPLAN results in that state (that is, Year 9 results can be used to pre-qualify for the minimum literacy and numeracy standards requirements for the Western Australian Certificate of Education).

17 This is due to the same logic noted in Anders et al. (2021), which found that students absent on the day of PISA testing are more likely to be lower achievers, as compared with the broader student population or students of similar characteristics in the population. A study of NSW government school students using NAPLAN data confirms this too applies for NAPLAN (CESE 2016). Though imputation can help to correct for this, it may not fully remove the bias from the missing data.

18 The spread across year levels is a product of the PISA sample being defined by age (15-year-olds) rather than year level as seen in the NAPLAN, PIRLS and TIMSS tests.

Equating introduces another source of uncertainty to the measurement process (known as equating error), which may affect the interpretability of performance trends.[19] For example, if one year's test difficulty is overestimated in the equating process, then results in that year, for all students and all student groups, would be overestimated. When the size of the equating error[20] is larger than that of the underlying year-on-year variation in the performance indicator being measured, it can become the dominating factor driving the trend.

Changes in scaling and equating processes over time may affect the results and trends of an assessment. ACER's preliminary investigation observes that there have been no substantive changes to the scaling and equating processes for the NAP assessments. But use of the same scaling and equating process does not mean the impact of equating error on results interpretation is consistent (e.g. biasing results in the same direction and by the same magnitude) across years. For example, the same NAPLAN equating process used in the past decade could mean 2021 test results being overestimated by 5 scaled points or 2022 results being underestimated by 15 scaled points. This could impact the trends.[21]

Assessment format

All four NAP assessments are transitioning to online testing, so the extent to which students are familiar with digital devices is becoming an increasingly important factor in interpreting results.

Despite concerns about different results from paper-based and online testing, the OECD's (2016) PISA field trial found that there were few countries where the mode of testing (that is, online or paper) caused a statistically significant effect on student performance. However, Jerrim et al. (2018) also examined PISA field trial results in Germany, Ireland and Sweden and found that on average students scored lower in computer-based assessments than in paper-based assessments. They tested the method that the OECD used to account for mode effects, which used questions that were thought to be equally difficult in both online and paper-based versions. They found that any effect caused by the mode of the test was likely to be small in mathematical literacy, but may have had impact in science, where the computer-based group still performed below the paper-based group in Ireland and Germany.

The mode effect warrants continued investigation, given the first assessment to shift online was PISA in 2015.[22]

Assessment design

All NAP assessments are also moving to use a form of adaptive testing to better measure student performance across the whole range of achievement, as highlighted in ACER's analysis for AERO. Adaptive assessments adjust the difficulty of the assessment to student performance, making the questions more challenging following correct answers or easier after incorrect answers. The level of adaptation varies in each assessment, and each is in a different phase of implementation. Since adaptive testing is a relatively new development, this again cannot explain the divergence in PISA and other assessment results that has been observed since the early 2000s.

Question design

NAPLAN, PIRLS, TIMSS and PISA all assess different aspects of literacy and numeracy, using different combinations of text types, lengths of texts and number of items per text.

For instance, reading load (or average words per question) differs between the assessments. In numeracy, NAPLAN questions tend to have simple contexts that are often abstracted to reduce reading load, as well as context-free problems with minimal reading. In TIMSS, about 85% of the numeracy

19 NAPLAN equating is further complicated as consideration is given to equating over year levels. TIMSS does not equate Year 4 and Year 8 results.

20 There are other types of random errors such as measurement error that contribute to the uncertainty of student performance in a given test; however, these tend to be smaller in size than equating error when we are looking at high-level performance trends (for example,. national and state and territory means) from assessments administered to large samples.

21 NAPLAN's equating process is different to the international assessments, which use common-item equating (that is, using a set of identical non-publicly released items). This is not possible with NAPLAN where it is important to make all test materials publicly available.

22 NAPLAN began shifting online in 2018, and all testing was online by 2022. Neither TIMSS nor PIRLS have been conducted online in Australia, though both are expected to be from 2023 on.

questions are situated in a problem-solving context (which range from straightforward to complex) (Mullis et al. 2021). In contrast, the numeracy questions for PISA tests are often heavily contextualised and usually contain a higher reading load than either NAPLAN or TIMSS items.

In literacy, the three assessments differ in the length of the stimulus texts provided. ACER's analysis for AERO found that PISA uses a wide range of text lengths, ranging from fewer than 100 words in a single text to lengthy and complex multi-screen digital texts, where part of the reading task is to retrieve relevant information via close reading. PIRLS (print) texts are typically 500 to 800 words in length, which is relatively lengthy considering the age of the tested cohort (Year 4 students). In contrast, NAPLAN texts are relatively short, with each year level set a maximum text length. This ranges from 250 words for Year 3 to 350 words for Year 9, which is much shorter than those typical of PIRLS.

Finally, PISA uses more scenario-based stimulus texts than the other three assessments, reflecting its overall objective to encourage the application of skills to real-world problem-solving.

ACER's preliminary analysis on behalf of AERO could not make a clear determination of whether these differences in question design might explain some of the divergence in PISA scores. It is possible that Australian students have become less familiar with scenario-based stimulus questions, or are increasingly finding high reading load questions challenging, due to a declining exposure to this type of question.

Content and curriculum coverage

The different content of the assessments and their relationship to the Australian Curriculum may play a part in explaining the different trends observed. NAPLAN content is aligned to the Australian Curriculum, while PIRLS, TIMSS and PISA, as international assessments, are not.

Previous research has established that the different content balance of the tests helps explain why countries may perform differently in different tests. Wu (2009) compared country-level results in TIMSS and PISA. While she found a high correlation between a country's result on each test, she concluded that where there were differences in results, these could largely be attributed to different content in the tests. Wu found that the differences in content balance of the tests (for example, with PISA having more data items and fewer algebra items), along with differences in the ages at which students took TIMSS (as a measure of how many years of schooling they would have experienced by the time they took PISA), collectively explained 93% of the variance of the differences in each country's performance in PISA and TIMSS. There has been little research into whether the concepts being tested in each assessment have been covered in the classroom by the time the test is administered. A 'test-curricula matching' exercise is done for TIMSS, but not PISA or PIRLS.

When curriculum matching was conducted for TIMSS 2019 Year 4, only 59 out of 171 items were expected to have been taught to Australian students by the end of Year 4. While this is a low proportion, if the assessment had been restricted to those 59 items, Australia's mean score would have increased only slightly – from 516 to 521. The comparable figures for Year 8 were 188 out of 206 items, with a possible score increase from 517 to 518.

Though it seems unlikely that a decrease in test-curricula matching in PISA would have occurred, the exercise could be attempted across each of the prior collection years, following the same process as the TIMSS exercise. This would clarify whether Australian students may have been less exposed to the type of skills and knowledge tested in PISA over time, either due to drift in what is assessed or what is taught.

Summary

The dramatic downturn in performance over time in PISA is not consistent with student performance trends in any of the other NAP assessments. Research into this divergence suggests that there is no simple explanation such as issues with assessment design, collection and reporting processes, or differences in question design and content and curriculum coverage. It may be a combination of all these issues, as well as a signal of true decline in student performance in the constructs measured by PISA over time. Further research may offer more insights into what diverging NAP trends really mean for Australian schooling.

3. Making use of assessment data to understand 'what works'

This section explores how the NAP assessments can be used to help us understand more about how effective our policies, practices and programs are, in raising student achievement.

Assessments can and should be used to make evidence-based decisions at all levels in education: from the classroom, to the whole school, to the whole system. Assessments facilitate policymaking when they provide relevant, accessible and quality information about student performance and the reasons for differences in student performance (Forster 2000). By providing high-quality data on factors that influence student achievements, assessments can also serve as a resource for identifying issues and developing education policy reforms (McGaw 2008; Wagemaker 2008).

This purpose of assessment is reflected in Australia's intergovernmental policy agreements. The Mparntwe Declaration (2019:19) sets out the need for good quality data to 'identify best practice and innovation' and 'develop a substantive evidence base on what works.' The National School Reform Agreement (2021) lists 'improving national data quality' as one of its 8 National Policy Initiatives, designed to enhance the national evidence base to inform policy development.

This section starts by examining the type of data required to make inferences about what works, before reviewing the limitations of using the NAP assessments for drawing policy inferences and suggesting some ways to address these limitations.

What assessment data can tell us

Student achievement data from the NAP assessments not only tell us whether young Australians are meeting important educational outcomes but can also provide information about the factors associated with higher achievement, which can inform policy and program design.

The international assessments collect additional information, beyond test responses, in the form of surveys of students and school staff. This information offers further insights that may support inferences about what works by allowing correlations between test scores and survey responses to be calculated (noting that the reliability of some of the existing surveys has been called into question, as discussed below). Using survey responses, we can potentially explore the impact of policies and practices; highlight practices that are used by high-performing countries; reveal changes over time; expose practices that are correlated with higher academic achievement; and identify practices that are promising candidates for further research. Given the international nature of these assessments, it is also possible to investigate differences in correlations across countries.

Questions about 'what works' may be explored at several levels. For example, in the classroom, research may consider elements within a teacher's locus of control, such as instructional practices. At the whole school level, factors such as leadership or school culture may be pertinent. At the system level, factors that are influenced by policy at the jurisdictional or national level – such as school funding – may be explored. The utility of the NAP assessments for making inferences about 'what works' across different levels of the education sector is considered below.

What data can support research at the classroom level?

The impact of classroom-level factors such as teaching practices is a frequent topic for exploration using international test and survey data. For example, studies have shown the following to be associated with higher levels of achievement:

- more frequent feedback on how to improve and current strengths (Grajcevci and Shala 2021)
- lower levels of classroom disruption (Blank and Shavit 2016)
- effective use of digital technologies (Petko et al. 2017).

This research has been made possible by the rich surveys of students and teachers collected as part of the international assessments, which give insights into classroom-level factors that administrative data alone cannot provide.

A perennial topic of interest at the classroom level is whether the use of specific teaching practices or pedagogies, such as inquiry-based teaching, influence student achievement (Kang and Keinonen 2018; Oliver et al. 2021). Inquiry-based teaching involves active learning by students, asking them to develop their own understanding of concepts and acquire knowledge through investigation, rather than directly from teachers (Jerrim et al. 2019). An influential study on this topic from McKinsey used PISA data to analyse the relationship between both inquiry-based teaching and teacher-directed instruction, and their influence on student achievement in science (Mourshed et al. 2017). Exposure to both teaching methods was measured using student survey data. The McKinsey study found that students achieved the best results when the 2 styles were used together to create a 'sweet spot', in which inquiry-based teaching was used in some lessons and teacher-directed instruction in many to all lessons. Oceania-specific analysis (there was no country-specific analysis undertaken) suggested the use of both styles at the sweet spot was associated with a 24-point increase in student scores, compared with their use in none to few lessons, while using inquiry-based methods in many to all lessons was associated with a 70-point decrease in student scores (Chen et al. 2017).

However, a more recent study that has incorporated a measure of prior achievement into the analysis has called this finding into question. Jerrim et al. (2019) studied the relationship between inquiry-based teaching and student achievement in science in England, linking PISA data to prior achievement measures from an externally marked examination at the end of primary school. They found little evidence that inquiry-based instruction is ever positively associated with students' academic achievement.

This example shows that the extensive student and staff surveys collected as part of the NAP international assessments can give useful insights into the prevalence of certain classroom practices, but they are insufficient for making reliable inferences about what works. Prior achievement data would add an important extra insight. This can be gained using data linkages and longitudinal studies (such as occurred in Australia with the 2015 PISA test data, which is linked to NAPLAN test data via the Longitudinal Surveys of Australian Youth).

What data can support research at the school level?

School-level factors have also been explored using international assessment data, with studies reporting the following factors as associated with higher levels of achievement:

- more frequent teacher collaboration (Mora-Ruano et al. 2019)
- positive school climate (Gómez and Suárez 2020)
- greater parental involvement (Sebastian et al. 2017).
- In particular, studies using PISA data have found evidence that principal leadership directly and indirectly impacts student achievement:
- Wu et al. (2020) used PISA teacher survey data relating to principals (asking teachers about the extent to which they agreed, for example, that the principal was aware of teachers' needs) to create an aggregate measure of principal effectiveness. They found that greater principal effectiveness was associated with student achievement in science. Contrary to their expectations, this relationship was not mediated by a principal's impact on school processes (such as increasing teachers' job satisfaction or collaboration).
- Tan (2018) examined principals' responses to questions of how frequently they used specific leadership behaviours to construct indices of different practices, such as instructional leadership, distributed leadership, goal-setting and problem-solving. They found that instructional leadership (for example, promoting teaching practices based on recent educational research) was positively associated with student achievement, with the greatest impact for disadvantaged students.

The international NAP assessments also provide survey responses about the management of staff, resources, technology, the school environment and its academic practices, which are needed to make reliable inferences on more abstract influences like school culture. Further investigation of NAP assessment data linked with administrative records, such as staff turnover or years of experience of school leaders, can enable deeper insights at the school level.

What data can support research at the system level?

At the system level, factors associated with greater student achievement include:

- fewer shortages of material resources (Hanushek and Woessmann 2017)
- greater school autonomy over hiring staff (Woessmann 2016)
- school choice and competition (Woessmann et al. 2007).
- In particular, PISA and TIMSS data have been used to make inferences about the efficacy of accountability mechanisms:
- Using PISA data, Woessmann et al. (2007) found that students perform better in systems where there is monitoring of student achievement (through external exit exams), monitoring of teacher practice (through observation of lessons) and monitoring of schools (through assessment-based comparisons). The combined impact of these practices amounted to a difference in student achievement of more than one and a half grade levels.
- Using TIMSS and PISA data, Woessmann (2005) found that students in countries with external examinations perform better than students from countries that do not have external examinations, with the difference in performance roughly the equivalent of one grade level and this impact being felt evenly across student groups regardless of family background. They also found that having external exams at the end of secondary school has a large impact on student achievement later in their schooling.

However, other studies have reported mixed results for other accountability mechanisms. Torres (2021) used PISA data to examine the impact of posting school achievement data publicly. They used 4 PISA cycles (2006–2015) to construct a measure of the proportion of students from each country who attend schools that post results publicly, as a proxy for how common this practice is in each country. For low- and middle-income countries, they found a positive association between accountability and student achievement in numeracy and science. However, for high-income countries, they found no relationship between accountability measures and educational outcomes in numeracy and science, and only a weak negative relationship between accountability and reading performance.

To support system-level inferences, student achievement data need to be linked to clear and comparable measures of policies and practices. The international NAP assessments provide this data on policies through surveys of school leaders and (in TIMSS and PIRLS) the national research coordinator from each country, with further information supplemented by other databases. NAPLAN data can be linked to the system-level settings of different jurisdictions in Australia to facilitate policy inferences. The NAP offers the possibility of international comparison with the range of policy options that exist outside of Australia.

The limitations of using assessment data to identify 'what works' – and possible solutions

Prior achievement

Previous research has established the difficulty of using assessment data to reach conclusions about the effectiveness of practices and policies that hold true for different subjects, year levels and contexts. A review by Deloitte Access Economics (2019) identified differences in question construction and interpretation, as well as a lack of data on moderating factors (for example, prior achievement), as barriers to understanding the relationship between practices and student achievement.

Lack of a prior achievement measure in the international NAP assessments limits their research utility. Without a prior achievement measure, analysis may be confounded if certain practices are more likely to be adopted for low or high achievers. This creates the risk of misrepresenting the true effect of practices on student performance.[23]

23 For example, without a prior achievement score, analysis could mistakenly find that providing more feedback to students is associated with lower levels of achievement, which may be a product of lower achievers receiving more feedback than their peers, rather than a reasonable assessment of the impact of giving feedback.

In a similar way, the learning environment documented in contextual survey questions may only partially reflect the earlier environment that has shaped students' achievement. The context questions act as an imperfect proxy for students' cumulative learning environments by focusing on their current school, which may underestimate the true impact of learning environment on achievement. This is particularly the case for the Year 8 TIMSS data, as any student not at a K–12 school will only have been in their school for a little over a year. This means that much of their academic development will have taken place earlier (in primary school), which may be quite different to the school described in the context survey.

Data linkage

In Australia the disconnection of assessment data from associated administrative data and contextual survey data places significant limits on the ability of educational research to deliver robust findings and to explore certain topics. More data linkage across these data sets is key to addressing this; a point that has been made by both Deloitte Access Economics (2019) and the Productivity Commission (2016).

Currently, NAP assessment data are only linked at an enduring national level to demographic and contextual data in certain longitudinal data sets. Each of the Longitudinal Study of Australian Children (LSAC), the Longitudinal Study of Indigenous Children (LSIC) and the Longitudinal Surveys of Australian Youth (LSAY) (2015 cohort alone) have linked survey-derived responses with NAPLAN records.[24] Additionally, the sampling for the 2003, 2006, 2009 and 2015 LSAY cohorts have been aligned to PISA test participation, which links results from that assessment to the survey-derived data. Uniquely, this means the 2015 LSAY cohort has a linkage to both PISA and NAPLAN data, which offers the possibility of researching performance across assessments. No linkages of TIMSS or PIRLS data exist.

The linked data sets have created useful resources for researchers to study associations of various factors with achievement over time. However, they have their limitations. Each covers only a small sample of their target cohort, and participant numbers reduce over time. More importantly, as they are periodic collections that do not enrol a new cohort each calendar year, they provide a limited number of data points over time compared with an annual collection.

At the system level, administrative records that include rich student demographic data, achievement data (from NAPLAN and other assessments) and school information (such as attendance data) do exist. The limitations of sample-derived records do not apply to these administrative records, but they face their own set of barriers to being a useful input into research and evaluation. As system-level records, they segment students by sector or by states and territories, which makes it impossible to draw insights that apply to all students within a state or territory, or across Australia. Further, each authority collects and links different measures as part of its administrative records. This impairs research and evaluation of both cross-sector and cross-jurisdiction initiatives.

A national student data set

Pooling administrative records to create an enduring integrated data set on a cross-sector and national basis would be transformational in addressing the limitations of both existing longitudinal studies and system-level administrative records. To be useful, such a data set would need to extend significantly beyond what is currently available from ACARA. At present, though student-level records for all NAPLAN participants can be provided by ACARA, the associated demographic information is not comprehensive,[25] and NAPLAN scores are presently only available in a form that links results from one test to the test prior (for example, a record with Year 9 NAPLAN scores will only include students' Year 7 NAPLAN scores).

24 For LSAC and LSIC, the NAPLAN linkage provides data that is contemporaneous to the survey-derived data, while for LSAY the NAPLAN linkage provides data covering the period before the first wave of surveying for the 2015 cohort.

25 At a student level, the only demographic data available are the students' gender, age, Aboriginal or Torres Strait Islander status, 'language background other than English' status, parents' highest level of educational achievement and occupational group, and the school's location and sector. For some indicators, data are missing or 'not stated', which further complicates any inferences drawn using this data.

A model of rich data is the National Pupil Database (NPD) in England. The NPD is a student-level administrative data resource curated by the UK government's Department for Education that has been found to be an extremely valuable resource for researchers, providing a near-complete picture of student trajectories and outcomes within the government sector.[26] It covers students from entry into the government-run early years system, through to when they exit school (it has also been linked to vocational and higher education study records to extend the utility of the data set in assessing post-school outcomes). The NPD includes details on all nationwide assessments undertaken throughout the early years and schooling, as well as rich demographic information, including language spoken at home, ethnicity and special education needs status. School exclusion and attendance records are also incorporated. It can be accessed by government and non-government researchers via an application process managed by the Department for Education.

The NPD has enabled interventions to be evaluated, relationships to be explored between disparate factors and has provided essential information on comparative effectiveness of reforms and initiatives (Jay et al. 2019).

Creating a cross-sector, Australia-wide version of the NPD would significantly improve the basis for making policy and program decisions, by enabling rigorous research and evaluation. It could initially draw on the data held by systems to enable linkage to additional data that will be collected into the future; for instance, forthcoming PISA, PIRLS and TIMSS records from the students these assessments sample (which would in turn enable trends across the NAP assessments to be better understood).

A national, cross-sectoral student data set may be one way forward to overcome limitations in the use of NAP data in policymaking and program design. By linking NAP achievement data with system-held demographic and school record data (for example, school absences) at a student level, more reliable cross-sector program and policy evaluations could be undertaken at a national level, as could robust exploratory research into drivers of educational outcomes. It could also enable research into cross-assessment trends by linking TIMSS, PIRLS and PISA with NAPLAN records.

Currently only a limited amount of demographic data can be accessed at a national level in a form linked to NAPLAN data, which are only available as an extract that spans the results of 2 test years (that is, there is no ability, at the national level, to track the progress of a student across Years 3, 5, 7 and 9). A national resource would also benefit systems by providing a detailed cross-sector view of student learning trajectories.

As the Productivity Commission (2016) noted, the main barriers to national cross-sector data linkage are privacy legislation that governs the use of personal information and a risk-averse culture among data custodians. However, with significant advances in systems for controlling the secure storage and use of record-level data (such as the development of the Australian Bureau of Statistics' Multi-Agency Data Integration Project[27]), these barriers stand a greater chance of being overcome.

Systematic surveys

Another limitation to the inferences that may be drawn from NAP assessment data derives from the nature of the survey questions in the international NAP assessments. The surveys may not fully capture variations in the quality, rigour or implementation of practices or policies. Questions about teaching practices tend to focus on the frequency of using certain practices and tools, rather than the quality of their use. Furthermore, the questions may be vaguely worded or lack consistency in how practices are identified, meaning that interpretation of student responses may be contestable. This generality is necessary to establish measurement equivalence across countries, however, it can lead to some researchers disagreeing about how to categorise teaching practices (for example, Jerrim et al. 2019).

Interpreting the information provided by students, teachers and principals is not straightforward, particularly in PISA. Respondents may interpret terms or phrases differently to each other or to how the survey developers intended (Rutkowski

26 The non-government sector in England educates about 6% of students, making it substantially smaller than Australia's in relative terms.

27 For more information on the Multi-Agency Data Integration Project (MADIP) see here.

and Rutkowski 2010), particularly when contrasted against a teacher's interpretation of how often certain practices are used in their classroom. In PISA, 15 year old students are surveyed at random within a school, rather than surveying whole classes like in PIRLS and TIMSS (Cordero Ferrera et al. 30 May 2016). This makes the process of interpreting teacher and student surveys more difficult, as the teacher questionnaire is given to teachers from different domains who teach 15-year-old students in the school, rather than those who teach the particular students who take the PISA assessment (Hillman and Thomson 2021). Principals are asked to provide information about their schools but may not be the most accurate source for some information related to teachers, such as teachers' morale and commitment.

Though the content of international assessment surveys is not in the control of any one country, further research into what questions to ask and how best to ask them to gain accurate insights into classroom and school approaches may show how improvements can be made. As the international assessments allow the insertion of a limited number of country-specific questions in student and teacher surveys, it would also be possible for improvements in question design to be unilaterally implemented.

Limitations of the international assessment surveys could also be overcome by collecting contextual information about the classroom and school from a representative sample of Australian students or teachers as part of NAPLAN. This could provide insights into a range of areas of interest, such as teacher practices; teacher understanding and use of assessment and assessment data; student experience of specific teaching practices; student wellbeing and sense of belonging; school environments; classroom management; and other factors that have previously been found to impact student achievement. Collecting such data through an additional survey could enable more informative educational research to be produced, with stronger insights into the relationship between student achievement and underlying factors of interest.

Such a survey would overcome current issues in using ad hoc surveys to make inferences about the prevalence of impact of certain practices. For example, to explore the practices used by high-performing schools, ACARA (2020b) identified schools that consistently made high progress in NAPLAN before inviting them to complete a survey on what practices they employ. While responses from schools show some consistency (for example, in using explicit instruction), the fact that the survey did not go out to a wider group of schools (that is, middle- and lower-performing schools) means concluding that these practices are a driver of high performance is not possible (given it is unclear whether they are equally prevalent in low-performing schools).

Collecting more and higher-quality data alongside the existing assessments will need to be carefully considered to ensure such a step would not have perverse consequences. For example, could the addition of a survey to NAPLAN lead to a further decline in participation? International examples may offer useful insights.

Similar surveys take place alongside national standardised testing in Japan and South Korea. Students are asked about their home and school environment, attitude to learning, after-school study and extracurricular activities (Morozumi and Tanaka 2020; Ra et al. 2019). In South Korea, school leaders are also asked about school demographics, the school climate, its use of practices such as ability grouping and how assessment results are used in the school (Ra et al. 2019).

Further research

There are fundamental limits on using observational and administrative data and the extent to which causal claims can be made from this data. As the OECD (2010:18) acknowledges, 'PISA cannot identify cause-and-effect relationships between inputs, processes and educational outcomes.' The OECD and International Association for the Evaluation of Educational Achievement (IEA) report on correlations between practices and student achievement, indicating the strength and direction of a relationship between two variables, but this does not establish causality between the two. Other factors may affect this relationship, such as whether the set of policies and practices were equally available to high-performing and lower-performing students, schools and systems. To understand the causal nature of the change, other forms of research and data collection need to occur (for example, randomised controlled trials).

Summary

The NAP suite of assessments and additional data could be used far more effectively to achieve their important purpose of monitoring and benchmarking student achievement, and helping us understand the effectiveness of our education practices, policies and programs. Current limitations can be addressed.

References

Commonwealth of Australia (2010) Question on Notice: 6, *Answers to Questions on Notice from the Australian Curriculum, Assessment and Reporting Authority*, The Senate, accessed 2 February 2022. https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Education_and_Employment/Naplan13/Additional_Documents

ACARA (Australian Curriculum, Assessment and Reporting Authority) (2016a) *NAP - Why NAP*, ACARA, accessed 23 December 2021. https://www.nap.edu.au/about/why-nap

ACARA (Australian Curriculum, Assessment and Reporting Authority) (2016b) *NAP - NAPLAN – general*, ACARA, accessed 1 February 2022. https://www.nap.edu.au/naplan/faqs/naplan--general

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

ACARA (Australian Curriculum, Assessment and Reporting Authority) (2020b) *What does it take to consistently deliver high progress in NAPLAN? [PDF]*, ACARA, accessed 21 December 2021. https://www.acara.edu.au/docs/default-source/media-releases/high-progress-schools-acara-media-release-20201214.pdf

ACARA (Australian Curriculum, Assessment and Reporting Authority) (2022) *NAPLAN National Report for 2022 [PDF],* ACARA. https://nap.edu.au/docs/default-source/default-document-library/2022-naplan-national-report.pdf

Ainley J, Cloney D and 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), accessed 23 March 2022. https://doi.org/10.1177/0004944120948654

Anders J, Has S, Jerrim J, Shure N and Zieger L (2021) 'Is Canada really an education superpower? The impact of non-participation on results from PISA 2015', *Educational Assessment, Evaluation and Accountability*, 33(1):229–249, doi:10.1007/s11092-020-09329-5.

Baker J and Cook H (28 June 2019) '"It's time": NSW wants NAPLAN replaced with "genuinely useful" test', *The Sydney Morning Herald*, accessed 1 February 2022. https://www.smh.com.au/education/it-s-time-nsw-wants-naplan-replaced-with-genuinely-useful-test-20190627-p5220m.html

Blank C and Shavit Y (2016) 'The association between student reports of classmates' disruptive behavior and student achievement', *AERA Open*, 2(3):2332858416653921, doi:10.1177/2332858416653921.

CESE (Centre for Education Statistics and Evaluation) (2016) *Mobility of students in NSW government schools [PDF],* NSW Department of Education. https://education.nsw.gov.au/content/dam/main-education/about-us/educational-data/cese/2016-mobility-of-students-in-nsw-government-schools.pdf

CESE (Centre for Education Statistics and Evaluation) (2019) I*dentifying potential strength and weakness in key learning areas using data from NAPLAN tests*, NSW Department of Education, accessed 25 January 2023. https://education.nsw.gov.au/about-us/educational-data/cese/publications/research-reports/identifying-strength-and-weakness-using-naplan-data.html

Chen L, Dorn E, Krawitz M, Lim C and Mourshed M (2017) *Drivers of student performance: Insights from Asia, McKinsey & Company [PDF],* McKinsey, , accessed 21 December 2021. https://www.mckinsey.com/~/media/mckinsey/industries/public%20and%20social%20sector/our%20insights/drivers%20of%20student%20performance%20asia%20insights%20revised/drivers-of-student-performance-insights-from-asia.pdf

COAG (Council of Australian Governments) (2018) *National School Reform Agreement, Department of Education and Training*, accessed 1 December 2021. https://www.dese.gov.au/quality-schools-package/resources/national-school-reform-agreement

Commonwealth of Australia (2010) *Administration and reporting of NAPLAN testing*, The Senate, accessed 1 February 2022. https://www.aph.gov.au/parliamentary_business/committees/senate/education_employment_and_workplace_relations/completed_inquiries/2010-13/naplan/report/index

Commonwealth of Australia (2014) *Effectiveness of the National Assessment Program – Literacy and Numeracy*, The Senate, accessed 2 February 2022. https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Education_and_Employment/Naplan13/Report/index

Cordero Ferrera JM, Gil Izquierdo M, and Universidad Autonoma de Madrid (30 May 2016) 'TALIS-PISA link: Guidelines for a robust quantitative analysis', in *6th Annual International Conference on Qualitative and Quantitative Economics Research (QQE 2016), Annual International Conference on Qualitative and Quantitative Economics Research (QQE 2016)*, Global Science & Technology Forum (GSTF), doi:10.5176/2251-2012_QQE16.19.

DAE (Deloitte Access Economics) (2019) *Unpacking drivers of learning outcomes of students from different backgrounds*, Department of Education, Skills and Employment, accessed 21 December 2021. https://www.dese.gov.au/integrated-data-research/resources/unpacking-drivers-learning-outcomes-students-different-backgrounds

Department of Education and Training Victoria (2021) *Equity (Catch Up) (Reference 12)*, Department of Education, accessed 23 December 2021. http://www2.education.vic.gov.au/pal/student-resource-package-srp-equity-funding-student-based-funding/guidance/2-equity-catch

Department of Education, Skills and Employment (2021) *Improving national data quality, Department of Education, Skills and Employment*, accessed 31 January 2023. https://www.education.gov.au/quality-schools-package/resources/improving-national-data-quality

Education Ministers (2019) *Alice Springs (Mparntwe) Education Declaration*, Department of Education, Skills and Employment. https://www.dese.gov.au/alice-springs-mparntwe-education-declaration/resources/alice-springs-mparntwe-education-declaration

Education Ministers (2022) *Education Ministers Meeting Communique - 16 March 2022*, Department of Education, Skills and Employment, accessed 13 April 2022. https://www.dese.gov.au/education-ministers-meeting/resources/education-ministers-meeting-communique-16-march-2022

Figlio D and Loeb S (2011) 'School Accountability', in *Handbook of the Economics of Education*, Elsevier, doi:10.1016/B978-0-444-53429-3.00008-9.

Forster MM (2000) *A Policy Maker's Guide to International Achievement Studies* [PDF], Australian Council for Educational Research. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1000&context=policy_makers_guides

Gómez RL and Suárez AM (2020) 'Do inquiry-based teaching and school climate influence science achievement and critical thinking? Evidence from PISA 2015', *International Journal of STEM Education*, 7, doi:10.1186/s40594-020-00240-5.

Grajcevci A and Shala A (2021) 'A review of Kosovo's 2015 PISA results: Analysing the impact of teacher characteristics in student achievement', *International Journal of Instruction*, 14(1):489–506.

Hanushek E and Woessmann L (2017) 'School resources and student achievement: A review of cross-country economic research', doi:10.1007/978-3-319-43473-5_8.

Hillman K and Thomson S (2021) *2018 Australian TALIS-PISA Link Report*, Australian Council for Educational Research, doi:10.37517/978-1-74286-628-4.

Jay MA, Grath-Lone LM and Gilbert R (2019) 'Data resource: The National Pupil Database (NPD)', *International Journal of Population Data Science*, 4(1), doi:10.23889/ijpds.v4i1.1101.

Jerrim J, Micklewright J, Heine J-H, Salzer C and McKeown C (2018) 'PISA 2015: How big is the "mode effect" and what has been done about it?', *Oxford Review of Education*, 44(4):476–493, doi:10.1080/03054985.2018.1430025.

Jerrim J, Oliver M and Sims S (2019) 'The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England', *Learning and Instruction*, 61:35–44, doi:10.1016/j.learninstruc.2018.12.004.

Kang J and Keinonen T (2018) 'The effect of student-centered approaches on students' interest and achievement in Science: Relevant topic-based, open and guided inquiry-based, and discussion-based approaches', *Research in Science Education*, 48(4):865–885, doi:10.1007/s11165-016-9590-2.

Kostogriz A and Doecke B (2011) 'Standards-based accountability: Reification, responsibility and the ethical subject', *Teaching Education*, 22(4):397–412, doi:10.1080/10476210.2011.587870.

MCEETYA (Ministerial Council on Education, Employment, Training and Youth Affairs) (2009) 'Assessing student achievement in Australia 2009' [PDF], MCEETYA. http://www.curriculum.edu.au/verve/_resources/NAP_2009-Assess_Stud_Achiev_Aust-Parent_Info_Brochure.pdf

McGaw B (2008) 'The role of the OECD in international comparative studies of achievement', *Assessment in Education: Principles, Policy & Practice*, 15(3):223–243, doi:10.1080/09695940802417384.

McGaw B, Louden W and Wyatt-Smith C (2020) *NAPLAN Review Final Report [PDF]*, NAPLAN Review. https://naplanreview.com.au/pdfs/2020_NAPLAN_review_final_report.pdf

Mora-Ruano JG, Heine J-H and Gebhardt M (2019) 'Does teacher collaboration improve student achievement? Analysis of the German PISA 2012 sample', *Frontiers in Education*, 4:85, doi:10.3389/feduc.2019.00085.

Morozumi A and Tanaka R (2020) 'Should school-level results of national assessments be made public?', *SSRN Electronic Journal*, doi:10.2139/ssrn.3648790.

Mourshed M, Krawitz M and Dorn E (2017) *How to improve student educational outcomes: New insights from data analytics*, McKinsey & Company, accessed 21 December 2021. https://www.mckinsey.com/industries/education/our-insights/how-to-improve-student-educational-outcomes-new-insights-from-data-analytics

Mullis IVS, Martin MO and von Davier M (2021) *Assessment Frameworks for the TIMSS 2023*, Boston College, TIMSS & PIRLS International Study Center website, accessed 31 January 2023. https://timssandpirls.bc.edu/timss2023/frameworks

NSW Department of Education (2022) *Check-in assessment*, accessed 23 February 2022. https://education.nsw.gov.au/teaching-and-learning/curriculum/literacy-and-numeracy/assessment-resources/check-in-assessment.html

OECD (Organisation for Economic Co-operation and Development) (2010) *PISA 2009 Results: What makes a school successful?: Resources, policies and practices (Volume IV)*, OECD, doi:10.1787/9789264091559-en.

OECD (Organisation for Economic Co-operation and Development) (2011) *OECD reviews of evaluation and assessment in education: Australia 2011*, OECD Publishing, doi:10.1787/9789264116672-en.

OECD (Organisation for Economic Co-operation and Development) (2016) *PISA 2015 results (Volume I): Excellence and equity in education*, OECD, accessed 22 March 2022. http://read.oecd.ilibrary.org/education/pisa-2015-results-volume-i_9789264266490-en

OFAI (Online Formative Assessment Initiative) (2020) *Home - Online Formative Assessment Initiative*, OFAI, accessed 2 February 2022. https://www.ofai.edu.au/

Oliver M, McConney A and Woods-McConney A (2021) 'The efficacy of inquiry-based instruction in Science: A comparative analysis of six countries using PISA 2015', *Research in Science Education*, 51(2):595–616, doi:10.1007/s11165-019-09901-0.

Pereira MC (2011) *An analysis of Portuguese students' performance in the OECD Programme for International Student Assessment (PISA) [PDF],* Banco de Portugal, Economics and Research Department. https://www.bportugal.pt/sites/default/files/anexos/papers/ab201111_e.pdf

Petko D, Cantieni A and Prasse D (2017) 'Perceived quality of educational technology matters: A secondary analysis of students' ICT use, ICT-related attitudes, and PISA 2012 test scores', *Journal of Educational Computing Research*, 54(8):1070–1091, doi:10.1177/0735633116649373.

Polesel J, Rice S and Dulfer N (2014) 'The impact of high-stakes testing on curriculum and pedagogy: A teacher perspective from Australia', *Journal of Education Policy*, 29(5):640–657, doi:10.1080/02680939.2013.865082.

Productivity Commission (2016) National education evidence base, Productivity Commission website, accessed 19 June 2022. https://www.pc.gov.au/inquiries/completed/education-evidence

Ra S, Kim S and Rhee K (2019) *Developing national student assessment systems for quality education: Lessons from the Republic of Korea*, Asian Development Bank, doi:10.22617/TCS190597-2.

Rogers SL, Barblett L and Robinson K (2018) 'Parent and teacher perceptions of NAPLAN in a sample of Independent schools in Western Australia', *The Australian Educational Researcher*, 45(4):493–513, doi:10.1007/s13384-018-0270-2.

Rutkowski L and Rutkowski D (2010) 'Getting it "better": The importance of improving background questionnaires in international large-scale assessment', *Journal of Curriculum Studies*, 42(3):411–430, doi:10.1080/00220272.2010.487546.

Sebastian J, Moon J-M and Cunningham M (2017) 'The relationship of school-based parental involvement with student achievement: A comparison of principal and parent survey reports from PISA 2012', *Educational Studies*, 43(2):123–146, doi:10.1080/03055698.2016.1248900.

Tan CY (2018) 'Examining school leadership effects on student achievement: The role of contextual challenges and constraints', *Cambridge Journal of Education*, 48(1):21–45, doi:10.1080/030576 4X.2016.1221885.

Torres R (2021) *Does test-based school accountability have an impact on student achievement and equity in education?: A panel approach using PISA, OECD*, doi:10.1787/0798600f-en.

VCAA (Victorian Curriculum and Assessment Authority) (2013) *Using NAPLAN data diagnostically - an introductory guide for classroom teachers [PDF],* VCAA, accessed 2 February 2022. https://www.vcaa.vic.edu.au/Documents/naplan/teachersguide-usingnaplandata.pdf

Wagemaker H (2008) 'Choices and trade-offs: Reply to McGaw', *Assessment in Education: Principles, Policy & Practice*, 15(3):267–278, doi:10.1080/09695940802417491.

Woessmann L (2005) 'The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA', *Education Economics*, 13(2):143–169, doi:10.1080/09645290500031165.

Woessmann L (2016) 'The Importance of School Systems: Evidence from International Differences in Student Achievement', *Journal of Economic Perspectives*, 30(3):3–32, doi:10.1257/jep.30.3.3.

Woessmann L, Lüdemann E, Schütz G and West MR (2007) 'School accountability, autonomy, choice, and the level of student achievement: International evidence from PISA 2003', *OECD Education Working Papers*, 13, OECD, doi:10.1787/246402531617.

Wu H, Shen J, Zhang Y and Zheng Y (2020) 'Examining the effect of principal leadership on student Science achievement', *International Journal of Science Education*, 42(6):1017–1039, doi:10.1080/09500693.2020.1747664.

Wu M (2009) 'A comparison of PISA and TIMSS 2003 achievement results in Mathematics', *PROSPECTS*, 39(1):33, doi:10.1007/s11125-009-9109-y.

Appendix A: Further detail on the National Assessment Program

NAPLAN

NAPLAN assesses literacy and numeracy skills aligned to the Australian Curriculum and is administered annually to students in Years 3, 5, 7 and 9 across sectors and jurisdictions, and tracks how a child is progressing over time.

Content

NAPLAN focuses on the two domains of literacy and numeracy. It broadly reflects aspects of literacy and numeracy within the curriculum in all jurisdictions and, more generally, the national curriculum. The types of test questions and test formats are chosen so that they are familiar to students and teachers across Australia. There are separate sections for reading, writing, spelling, grammar and punctuation, and numeracy.

Format

The test is broken down into 4 sections: reading, writing, language conventions and numeracy, and takes between 170 and 230 minutes to complete, depending on the grade level. NAPLAN was originally offered as a paper-based test but began a shift to being offered online in 2018. All tests were administered online by 2022. Unlike the paper-based test, NAPLAN online tests have an adaptive test design, meaning how students respond to one question (or a set of questions) determines which question (or set of questions) they receive next. For example, if a student gets a question of medium difficulty correct, they might be shown a more challenging question next. If they get a question of medium difficulty incorrect, they might be shown an easy question next.

Administration

The test is offered in May and is managed by ACARA. Annual reports show test results broken down by student subgroups and in some cases show trends over time. It has been administered every year since 2008, except for 2020 when it was not offered due to the COVID-19 pandemic.

Sampling

NAPLAN is taken by all children in Years 3, 5, 7 and 9 across sectors and jurisdictions, subject to some exclusions.

This information comes from the NAPLAN National Report.

More information can be found in the 2022 NAPLAN National Report.

PISA

PISA is an international assessment of 15-year-olds' ability to apply their knowledge and skills to real-life problems and situations.

Content

The assessment focuses on three core academic domains: reading, mathematics and science. There are additional domains that countries can opt into.[28] Unique to PISA's approach is that it seeks not only to assess basic skills and ability to reproduce knowledge, but also to extrapolate knowledge to new situations. This approach sets PISA apart from other standardised tests.

In addition to the cognitive assessments, students also complete a student questionnaire about their family background, aspects of their lives such as their motivation and engagement towards learning, and their attitudes to school. Students are given up to an hour to complete the background questionnaire. School principals also complete a short web-based questionnaire that focuses on information about the level of resources in the school, the school environment and the qualifications of staff.[29]
The major domain of the cognitive assessment is also the focus of some of the questions in the student and school questionnaire. For instance, if the major domain is reading, students may be asked about their experience, motivations and attitudes related to reading. Schools may be asked to report about teaching practices for reading, teacher qualifications for those teaching reading, and school environmental factors and policies related to reading.

28 In 2018, PISA administered a test on an 'innovative domain' called Global Competency, which Australia did not participate in. It also administered a test on financial literacy that Australia did participate in.

29 PISA does also offer optional context questionnaires for parents and teachers, but Australia did not participate in these in 2018.

Format

The cognitive test is computer-based and takes 2 hours. PISA selects a major domain each year it administers the test. All participating students take the assessment in the major domain and take one additional cognitive assessment in one of the 2 remaining domains, determined on a randomised basis. For example, in a year when reading is the major domain, all students will take the reading assessment. In addition, 50% of students will take the mathematics assessment and 50% will take the science assessment. This means that 50% of test-taking time is spent on the major domain.

Administration

PISA is directed by the OECD, managed in Australia by ACER, and jointly funded by the Australian Government and all jurisdiction governments. PISA has been offered every 3 years since 2000. It is administered between July and September of the testing year. Results are reported on a 0–1000 scale with a mean of 500 and standard deviation of 100, set in the baseline year that the test was offered so that achievement trends can be measured over time. The scale for mathematics was changed in 2003 and the scale for science changed in 2006, so these are the baseline years for each of these tests.

Sampling

PISA uses a two-stage stratified sample. The first stage involves the sampling of schools in which 15-year-old students could be enrolled. The second stage of the selection process involves randomly sampling students within the sampled schools. The following variables are used in the stratification of the school sample: jurisdiction; school sector; geographic location; sex of students at the school; and a socioeconomic background variable (based on the Australian Bureau of Statistics' Socio-Economic Indexes for Areas, which consists of four indexes that rank geographic areas across Australia in terms of their relative socioeconomic advantage and disadvantage).

In 2018, the final sample included 17,448 students in 734 schools. This represented a weighted response rate of 95.8% after replacements and the weighted student participation rate after replacements was 85.1%. Both these figures met the international standards on response rates as specified by the Technical Advisory Group.

This information comes from PISA Student Performance Reports.

More information can be found in PISA 2018: Reporting Australia's Results. Volume I Student Performance.

TIMSS

TIMSS is an international study of mathematics and science achievement administered every 4 years since 1995 to Year 4 and Year 8 students.

Content

TIMSS is designed to align with the Year 4 and Year 8 mathematics and science curricula used in the participating education systems and countries. It is organised around a content dimension, which specifies the domains or subject matter to be assessed in mathematics and science, and a cognitive dimension, which specifies the thinking processes and sets of behaviours expected of students as they engage with the content.

In Year 4, there are three content domains in mathematics – number, measurement and geometry, and data – and three in science – life science, physical science and Earth science. In Year 8, there are four content domains in mathematics – number, algebra, geometry, and data and probability – and four in science – biology, chemistry, physics and Earth science. At both year levels, there are three cognitive domains in each curriculum area: knowing, applying and reasoning.

In addition to cognitive assessments of mathematics and science skills, TIMSS also provides data about students' contexts for learning mathematics and science based on questionnaires completed by students and their parents, teachers and school principals. The student questionnaire includes questions about home contexts, students' characteristics, and attitudes towards learning mathematics and science. Teacher and school questionnaires are also administered to the mathematics and science teacher(s) of each selected class and to the principal of the school. The teacher questionnaire includes questions about teacher

preparation and experience, pedagogical practices, use of technology, assessment, assignment of homework, school and classroom climate, and whether the TIMSS topics have been covered in class. The school questionnaire, answered by the principal, seeks descriptive information about school characteristics, instructional time, resources and technology, school climate for learning, students' school readiness, and principal preparation and experience.

Format

TIMSS is offered as a 72–90 minute paper-based assessment, with additional time provided for the background questionnaire.[30] For Year 4 students, the assessment is broken up into two 36-minute sessions with equal amounts of mathematics and science questions for each participating student. For Year 8 students, the assessment is broken up into two 45-minute sessions.

Administration

TIMSS is directed by the International Association for the Evaluation of Educational Achievement (IEA), managed in Australia by ACER, and jointly funded by the Australian Government and all jurisdictions. TIMSS has been offered every 4 years since 1995, except for in 1999. It is administered towards the end of the school year, between October and December of the testing year. Results are reported on a 0–1000 scale with a mean of 500 and standard deviation of 100, set in the baseline year of 1995 that the test was offered so that achievement trends can be measured over time.

Sampling

TIMSS uses a two-stage stratified sample design. In the first stage, schools are randomly sampled, stratified by jurisdiction, sector, geographic location and a socioeconomic variable to ensure national representation. In the second stage, one or two Year 4 or Year 8 classrooms in each selected school is randomly selected. Their principals and mathematics and science teachers are also asked to complete a survey.

In Australia, 287 primary schools and 284 secondary schools participated in the data collection for TIMSS 2019. From each school at least one intact class from the relevant year level was selected, resulting in a sample of 5890 Year 4 students and 9060 Year 8 students. Australia took a larger sample than the one required by TIMSS to ensure that reliable estimates could be inferred for the states and territories.

This information comes from TIMSS in Australia Student Performance Reports.

More information can be found in TIMSS 2019 Australia. Volume 1: Student performance.

PIRLS

PIRLS is an international study of Year 4 reading literacy achievement administered every 5 years starting in 2001.

Content

PIRLS is organised around 2 dimensions – the purposes for reading, and the processes of comprehension that readers use in understanding the texts and their related questions. The 2 purposes for reading in PIRLS are reading for literary experience and reading to acquire and use information. Within each of these 2 major reading purposes, 4 processes of comprehension are also assessed: (i) focusing on and retrieving explicitly stated information, (ii) making straightforward inferences, (iii) interpreting and integrating ideas and information, and (iv) examining and evaluating content, language, and textual elements. Overall, half of the PIRLS assessment focuses on reading for literary experience and half on reading to acquire and use information.

PIRLS also provides data about students' contexts for learning based on questionnaires completed by students and their parents, teachers and school principals, and curriculum experts from each country. Students complete a questionnaire about their home contexts, characteristics, and attitudes towards learning and reading.

30 In 2019, TIMSS began the transition to computer-based assessment by introducing a computerised version known as eTIMSS to half of the participating countries. Australia has not yet administered TIMSS as a computer-based assessment.

Reading/English teachers of student participants also complete a questionnaire about their teacher preparation and experience, pedagogical practices, use of technology, assessment, assignment of homework, school and classroom climate, and their own attitudes towards reading. Principals complete a school questionnaire, providing information about school characteristics, instructional time, resources and technology, school climate for learning, students' school readiness, and principal preparation and experience. ACER has submitted responses to a questionnaire about the reading curriculum, school organisational approaches and instructional practices in Australia. This response was reviewed by curriculum experts in each state and territory education department, and then submitted to the International Study Centre.

Format

PIRLS is offered as an 80-minute paper-based assessment, with additional time provided for the background questionnaire.

Teachers, principals and curriculum experts also completed questionnaires, which enabled collection of information about the various contexts of teaching and learning reading.

Administration

PIRLS is directed by the International Association for the Evaluation of Educational Achievement (IEA), managed in Australia by ACER, and jointly funded by the Australian Government and all jurisdictions. PIRLS has been offered internationally every 5 years since 2001 but Australia has only taken part in the past 3 rounds, in 2011, 2016 and 2021. It is administered towards the end of the school year, between October and December of the testing year. Results are reported on a 0–1000 scale with a mean of 500 and standard deviation of 100, set in the baseline year of 2001 so that achievement trends can be measured over time.

Sampling

PIRLS uses a two-stage stratified sample design. In the first stage, schools are randomly sampled, stratified by jurisdiction, sector, geographic location and a socioeconomic category for the area of each school to ensure national representation.In the second stage, one or two classrooms in each selected school is randomly selected. At each sampled school, one intact Year 4 class – along with all Aboriginal and Torres Strait Islander students in that year level – was selected to participate in PIRLS 2016. This resulted in a sample of 6341 Year 4 students. Results from PIRLS 2021 will be released in May 2023.

This information comes from PIRLS in Australia Reports.

More information can be found in PIRLS 2016: Reporting Australia's results.

Appendix B: Preliminary analysis conducted by the Australian Council for Educational Research (ACER)

In 2022, the Australian Education Research Organisation commissioned ACER to undertake preliminary analysis into the Programme for International Student Assessment (PISA); National Assessment Program – Literacy and Numeracy (NAPLAN); Trends in International Mathematics and Science Study (TIMSS); and Progress in International Reading Literacy Study (PIRLS).

This analysis covered a range of aspects of all 4 assessment programs – from curriculum coverage to student selection to implementation and data reporting –to identify the importance of each of these in interpreting the data generated by each one.

The analysis included a discussion of what each assessment aims to measure (for example, within a domain such as numeracy, the elements or skills that are focused on and the alignment of the assessment to the curriculum taught in schools), as well as describing how any differences in what assessments aim to (or actually) measure should influence interpretation of the data they provide.

The analysis also included the extent to which the design and collection of each measure should influence the interpretation of the information they provide, both individually and collectively. This included a discussion on the design of each assessment, and detailed the implications of design elements or collection processes (for example, sampling processes, response or participation rates, medium of assessment, measurement error, equating processes) for interpreting trends at a national and subgroup (for example, Aboriginal and Torres Strait Islander students) level.

For further information about this commissioned preliminary analysis, please contact AERO.

For more information visit
edresearch.edu.au

280223.RK.AA.V1