Au[redacted]

Contents

Glossary of terms and abbreviations

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

I t t

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. Rat a a Nat a A t P a (NAP)

I t t

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.

C t t NAP

assessment mix is able to meet emergis. [illegible]

Section 2

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.

T at a t t

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

[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).

D t NAP a t ?

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

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.

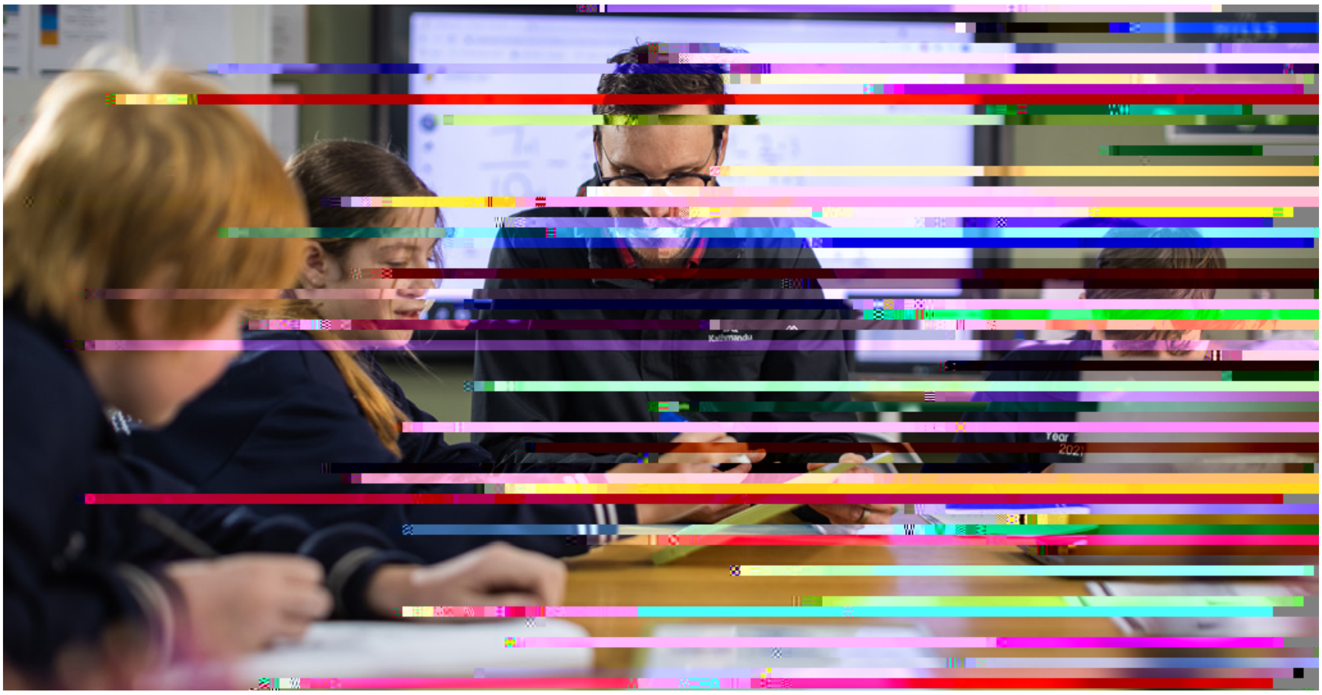
Upcoming changes to NAPLAN may provide an opportunity for teachers to better use NAPLAN data in a formative or diagnostic way. The education ministers have agreed that, from 2023, NAPLAN should be conducted earlier in the school year (in Term 1). This change, and the recent move to universal online delivery, 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.

S a

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.



2. W at at a a t t t

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.

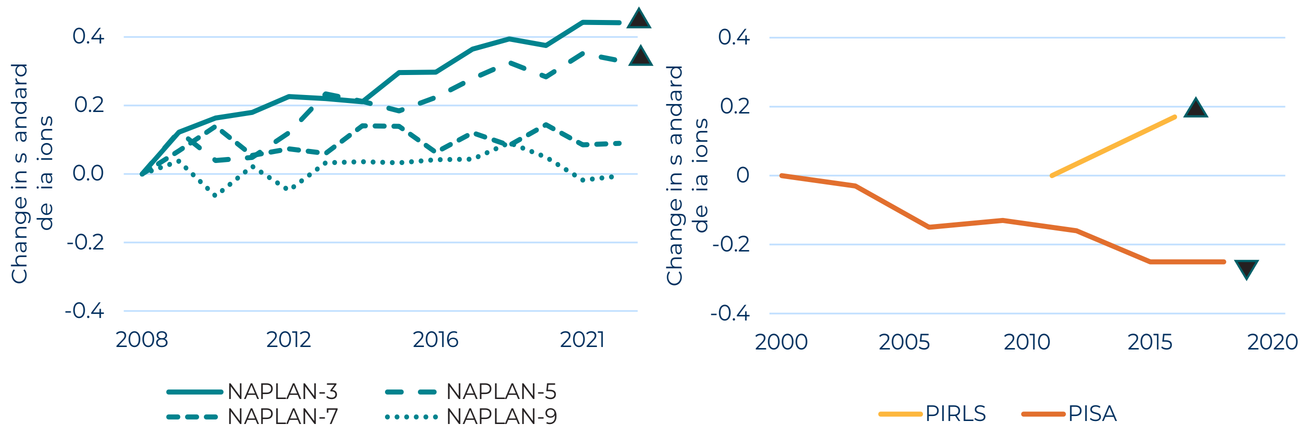
This section explores the disparate trends across the NAP assessments before considering whether the apparent divergence is unique to Australia. A preliminary investigation by the Australian Council for Educational Research (ACER), on behalf of AERO, considered possible explanations for the different 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.

D t t t NAP a t

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.

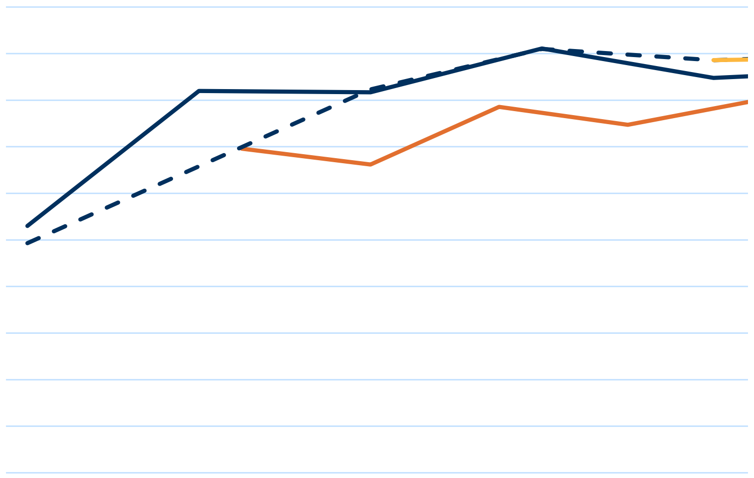
Figure 1 and Figure 2 provides a visual account of these trends.[9] To compare trends across these assessments, AERO calculated the change in average achievement for each assessment between a given year and the baseline year. This change is measured in standard deviation units (a measure of variation)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

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

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.

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.

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.

S a

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.

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

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.

[illegible]

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.

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 factoru2cvgsa 51.0236 123.6853 cm0 0 m70.866 0 lTi̇mygrhA0.29

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 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

R

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/

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_

A[illegible] A: F[illegible]t[illegible] [illegible]ta [illegible] t[illegible] Nat[illegible]a A[illegible]t P[illegible]a

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

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

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.mathematiple design.

en0.068 kQ0.75 w /P /Lan(gtics and science teachers are also 61g)TjEMC0 ng (enkssmen9)Clo-stage str /P 2019gtics and scie

Appendix B: Preliminary analysis commissioned from 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.

Australian
Education
Research
Organisation

For more information visit
edresearch.edu.au