GRATTAN Institute

Widening gaps

What NAPLAN tells us about student progress

Technical Report

Peter Goss and Cameron Chisholm

Grattan Institute Support

Founding members

Australian Government

VICTORIA State Government

THE UNIVERSITY OF MELBOURNE

bhpbilliton

Program support

Higher Education

THE MYER FOUNDATION

Affiliate Partners

Google

Origin Foundation

Medibank Private

Senior Affiliates

EY

PwC

The Scanlon Foundation

Wesfarmers

Affiliates

Ashurst

Corrs

Deloitte

GE ANZ

Urbis

Westpac

Grattan Institute Technical Report, March 2016

This technical report was written by Dr Peter Goss, Grattan Institute School Education Program Director, and Dr Cameron Chisholm, Grattan Institute Senior Associate. It was prepared to accompany the Grattan Institute Report, *Widening gaps: What NAPLAN tells us about student progress*. The purpose is to present the data and methodology used in the analysis, with a discussion exploring robustness and sensitivity.

We would like to thank the members of Grattan Institute's School Education Reference Group for their helpful comments on the methodology, as well as numerous industry participants and officials for their input. We would also like to thank two experts who specifically reviewed this technical report.

The opinions in the technical report are those of the authors and do not necessarily represent the views of Grattan Institute's founding members, affiliates, individual board members reference group members or reviewers. Any remaining errors or omissions are the responsibility of the authors.

Grattan Institute is an independent think-tank focused on Australian public policy. Our work is independent, practical and rigorous. We aim to improve policy outcomes by engaging with both decision-makers and the community.

For further information on the Institute's programs, or to join our mailing list, please go to: http://www.grattan.edu.au/

The technical report may be cited as:
Goss, P., and Chisholm, C., 2016, *Widening gaps: what NAPLAN tells us about student progress. Technical Report*, Grattan Institute

Overview

The report for Grattan Institute *Widening gaps: What NAPLAN tells us about student progress* seeks to measure and compare relative student progress on the National Assessment Program – Literacy and Numeracy (NAPLAN) test in a way that is robust, easy to interpret, and comparable across different groups of students. It analyses student-level data to identify some of the factors associated with higher or lower rates of progress, and to quantify the degree of these associations. The analysis does not attempt to quantify the causal impact of these factors, and should not be interpreted as such.

Every year since 2008, the NAPLAN test has been administered Australia-wide to nearly all students in Years 3, 5, 7, and 9. This means that students who were in Year 3 in either 2008 or 2009 have now taken the NAPLAN test across each of the test-taking years. This makes it possible to track how much students have progressed (as measured by NAPLAN) over a significant proportion of their time spent at school.

This technical report includes four technical appendices to *Widening gaps*. Appendix A describes the rationale and conceptual framework behind creating a new frame of reference to interpret NAPLAN results. Appendix B describes the data used in the analysis, and discusses some of the data issues. Appendix C outlines the technical detail behind the methodology to convert NAPLAN scale scores to *equivalent year levels*. Finally, Appendix D explains the approach used to track the progress of students from Year 3 to Year 9, using Victorian linked data.

Table of contents

List of Figures

A Conceptual framework for translating NAPLAN scale scores into equivalent year levels

A.1 The design of NAPLAN

A.1.1 NAPLAN scale scores

Students that undertake the NAPLAN test receive a score for each assessment domain: reading, writing, language conventions (which includes spelling, grammar and punctuation), and numeracy. This score, called the NAPLAN scale score, is typically between 0 and 1000. While the scores are used to indicate whether a student is above NAPLAN national minimum standards for each year level, they have no other direct interpretation. The scores are an estimate of student skill level at a point in time, a latent concept – the numbers themselves have no particular meaning.[1] Nor are the scores comparable across assessment domains.

A.1.2 Horizontal and vertical equating

The NAPLAN test is designed so that results in each domain can be compared between students in different year levels and students taking the test in different years. This means, for example, that a student who took the Year 5 NAPLAN reading test in 2012 and received a scale score of 500 is estimated to be at the equivalent level of a student who took the Year 7 reading test in 2013 and received the same score. That is, they are demonstrating comparable reading skills in the elements being tested by NAPLAN. This property of NAPLAN is achieved via a process known as *horizontal* and *vertical equating*.

The horizontal equating process involves a sample of students taking an equating test in addition to the NAPLAN tests. A scaling process takes place using this equating sample and common items across years on the equating tests. The result is that NAPLAN scale scores are comparable across different years. The vertical equating process involves common test items on the tests administered to different year levels. The results are scaled so that scale scores are comparable across different year levels.[2]

While the horizontal and vertical equating process is necessary to measure student progress over time, it also introduces an additional source of error into NAPLAN results.[3] The results presented in this analysis take the equating process as given, which means any errors arising from this process reduce the reliability of the analysis. We suggest that our analysis should be revisited after NAPLAN is moved online from 2017, as online testing is likely to strengthen the equating process.[4]

[1] It would be possible to link NAPLAN scale scores to curriculum standards, but this has not yet been developed. It is possible that NAPLAN scores will become more closely linked to curriculum standards with the move to NAPLAN online.

[2] See ACARA (2015e), pp. 40–72 for details.

[3] See, for instance, Wu (2010).

[4] ACARA (2015c) and Wu (2010).

A.2 Looking at progress through a new lens

A.2.1 NAPLAN scale scores give an incomplete picture of student progress

Student performance on standardised tests can be measured in a number of different ways.[5] The simplest measure, raw test scores, can be used to rank students. But raw scores can be hard to interpret. For example, on a 40-question test, the difference in skill level between a student with 25 correct answers and another with 20 correct answers should not be considered equal to the difference between a student with 40 correct answers and another with 35 correct answers. Raw test scores are even less useful for looking at student progress over time, because the measure does not take into account the degree of difficulty in the questions asked in different tests.

NAPLAN scale scores are developed from the Rasch model, an advanced psychometric model for estimating a student's skill level. The resulting estimates have a number of desirable properties, including being on an interval scale.[6] This property suggests that student progress can be measured by 'gain scores': the difference between NAPLAN scale scores in two test-taking years.[7] But there are limitations to using this measure, as ACARA notes:

> *It is important to consider that students generally show greater gains in literacy and numeracy in the earlier years than in the later years of schooling, and that students who start with lower NAPLAN scores tend to make greater gains over time than those who start with higher NAPLAN scores.*[8]

That is, the "path of progress" that students take across the four NAPLAN test years is not a linear function of the NAPLAN scale score, as shown in Figure A.1. Between 2012 and 2014 in numeracy, for instance, the median student made a gain of 86 points between Years 3 and 5 (an average of 43 points each year), 54 points between Years 5 and 7 (an average of 27 points each year), and 43 points between Years 7 and 9 (an average of 21.5 points each year).[9]

ACARA implicitly acknowledges this non-linear growth path in the way that NAPLAN proficiency bands are defined. Specifically, the national minimum standard jumps by two bands from Year 3 to Year 5, but only one band from Year 5 to Year 7 and from Year 7 to Year 9 (see Box A.1 on page 9). Even so, proficiency bands do not adequately take non-linearity into account and so we do not use them in the *Widening gaps* report.

Given that the observed growth in NAPLAN scores is not-linear with student year level, what does this mean? One interpretation would be to say that the education system is less effective for students in later year levels, especially between Year 7 and Year 9. This would be an important finding.

Of course, it could be that the smaller gain scores observed between higher year levels can be attributed to teaching differences – for instance, a shift from skill development to

[5] Angoff (1984).

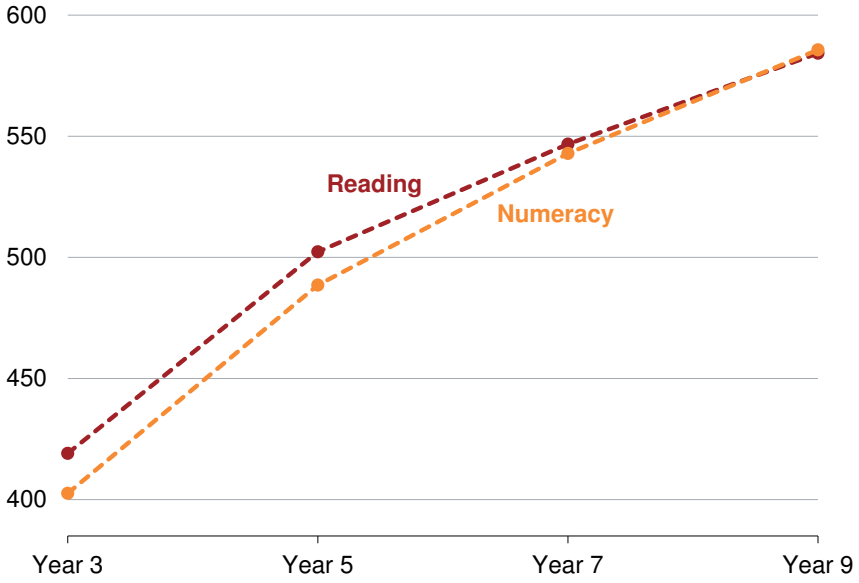
[6] This means that, in terms of skill level on the construct being tested, the difference between a score of 400 and 450 is equivalent to the difference between 600 and 650, for example.

[7] NAPLAN is a test of specific literacy and numeracy skills. These skills are fundamental to student learning. Yet a standardised test does not cover all elements of student learning; for instance, NAPLAN tends to focus on specific skills rather than content knowledge. Thus, when the report refers to 'learning' or 'progress' in numeracy or reading, it is referring to that which can be measured by NAPLAN.

[8] ACARA (2015b).

[9] Grattan analysis of ACARA (2014).

Figure A.1: The relationship between NAPLAN scale scores and year level is not linear for the median student
NAPLAN scale score of median student in each year level, Australia



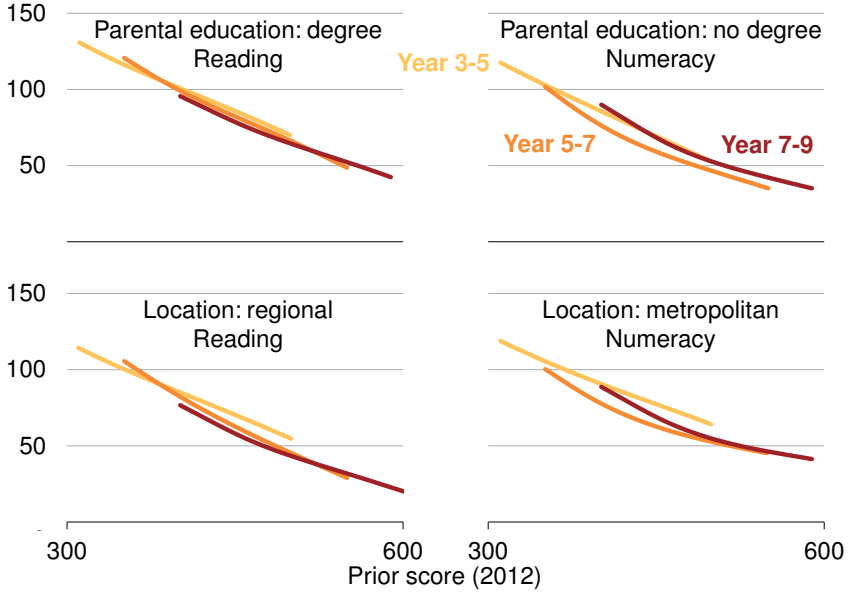
Notes: Based on 2014 and 2012 median scores.
Source: Grattan analysis of ACARA (2014).

content knowledge in secondary school. But if this was the case, we would expect gain scores to be strongly related to year level, and only weakly related to prior test score once year level is taken into account. Figure A.2 suggests that this is not the case: lower prior scores are associated with higher gain scores *within* each year level, and the same pattern holds for different population sub-groups.[10]

A third interpretation is that students genuinely increase their skill level faster from a lower base, and slow down over time. That is, the higher a student's current skill level, the longer it takes to

[10] Year level appears to have some effect, particularly for numeracy, but the impact is relatively weak once prior scores are taken into account.

Figure A.2: Higher gain scores are observed for lower prior scores, regardless of year level or population sub-group
Median NAPLAN gain score over two years by prior score, 2014, Australia



Notes: Similar patterns exist for other sub-groups, and at different percentiles. Gain scores estimated by a median quantile regression with cubic regression splines.
Source: Grattan analysis of ACARA (2014).

increase their skill level by a given amount (as measured by the NAPLAN scale). This appears to be the favoured interpretation among psychometricians.

Regardless of the explanation, this pattern of higher gain scores from lower starting scores should be taken into account when comparing the relative progress of different sub-groups of students. If not, it is too easy to draw spurious conclusions about the progress of different groups by over-interpreting gaps or gains in NAPLAN scores to mean something about broader learning progress.

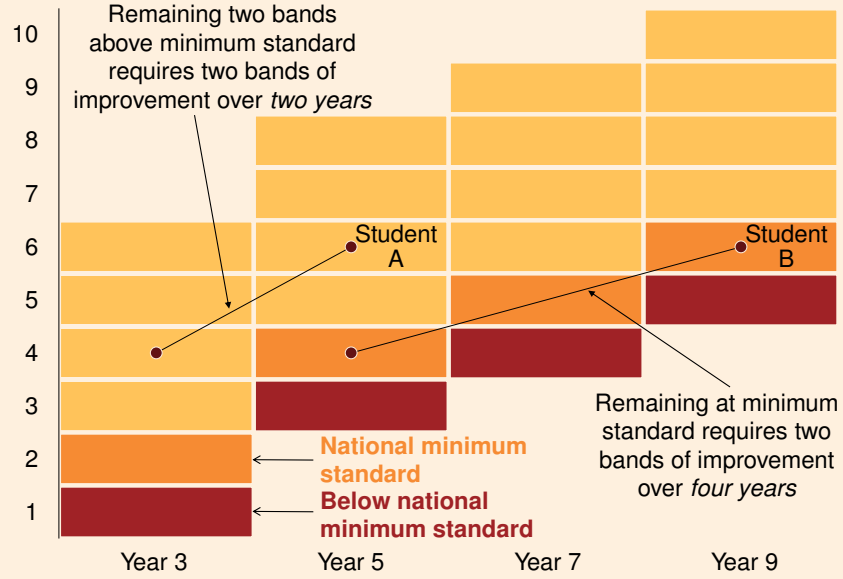
Box A.1: NAPLAN proficiency bands do not adequately take non-linearity into account

ACARA implicitly acknowledge the non-linear path of progress in the way that results are reported against *NAPLAN proficiency bands*. There are ten proficiency bands spanning Year 3 to Year 9, with equally-spaced cut-points along the NAPLAN scale.[a] These bands are used to define the National Minimum Standards. But because student skill level does not increase linearly over time, the National Minimum Standard increases by two bands between Years 3 and 5, but by only one band between Years 5 and 7 and between Years 7 and 9.[b] If there was reason to believe that the path of progress should be linear, then the change in the National Minimum Standards between each year level should be consistent.

Six proficiency bands are reported for each year level. For a student to remain in the same *relative* proficiency band, they must move up two bands between Years 3 and 5, then one band between Years 5 and 7, and another band between Years 7 and 9. But students who remain in the same relative band have not necessarily been progressing at the same rate.

Figure A.3 provides an example of this – Student A moves from Band 4 in Year 3 to Band 6 in Year 5, staying two bands above the national minimum standard. Student B performs consistently in the national minimum standard band, moving from Band 4 in Year 5 to Band 6 in Year 9. Both students remain in the same relative proficiency band, which suggests they are learning at the same rate. Yet Student A makes the same gain over two years as Student B does over four. This suggests that the non-linear scale of proficiency bands does not consistently account for the non-linear path of progress for students at different skill levels.[c]

Figure A.3: The level of growth required to remain in the same relative proficiency band changes with year level



Source: ACARA (2015e).

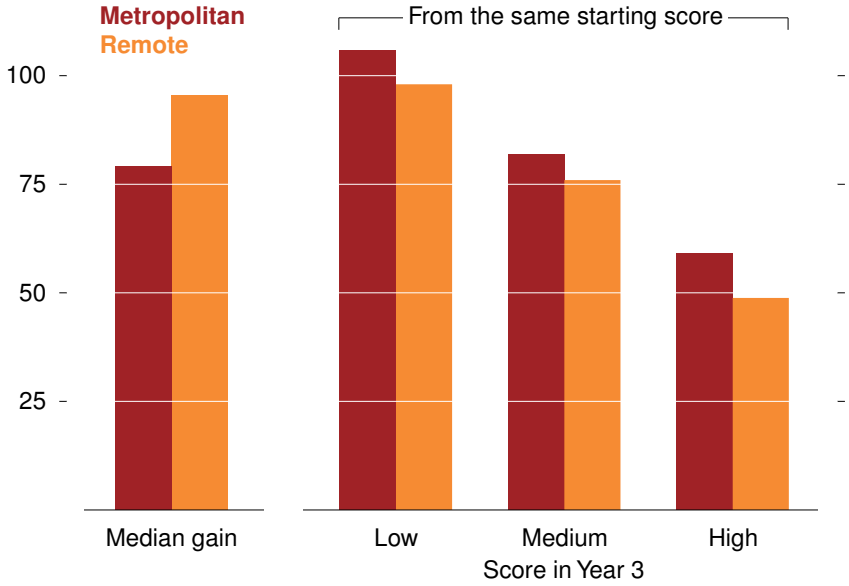
[a] With the exception of Band 1 and Band 10, each band spans 52 NAPLAN scale points.
[b] The National Minimum Standard is Band 2 for Year 3, Band 4 for Year 5, Band 5 for Year 7, and Band 6 for Year 9.
[c] The analysis in this report does not use NAPLAN proficiency bands to assess student progress.

For example, students from remote areas score below students from metropolitan areas in Year 3, yet make higher gain scores, on average.[11] That is, remote children are increasing their skill level, as measured by NAPLAN, by more than metropolitan children. But it would be incorrect to infer from this that the remote students are catching up to metropolitan students in a broader sense. To catch up, a student who is behind must at some stage learn faster (comparing the rates of learning over the same set of skills). In fact, when we compare the gain scores of remote and metropolitan students *from the same score in Year 3*, students from remote areas consistently make lower gain scores (at the median) than those from metropolitan areas, as shown for reading in Figure A.4. Remote students are actually falling further behind metropolitan students.

Some researchers have accounted for the non-linearity in the student path of progress using 'like-for-like' comparisons. That is, they have only compared gain scores across different sub-groups from the same prior score.[12] Like-for-like comparisons can be useful for interpreting gains made by different sub-groups, as the example with metropolitan and remote students shows. But this approach is limited in its scope – many population sub-groups start from very different skill levels. To compare the relative progress of students starting from different skill levels requires a new lens.

Figure A.4: Remote students make higher gains on average than metropolitan students, but lower gains from the same starting score

Median NAPLAN gain score between Year 3 (2012) and Year 5 (2014), reading, Australia



Notes: 'Low, medium and high' Year 3 scores are defined as the 20th, 50th, and 80th percentiles respectively. A similar pattern between metropolitan and remote students exists for numeracy.
Source: Grattan analysis of ACARA (2014).

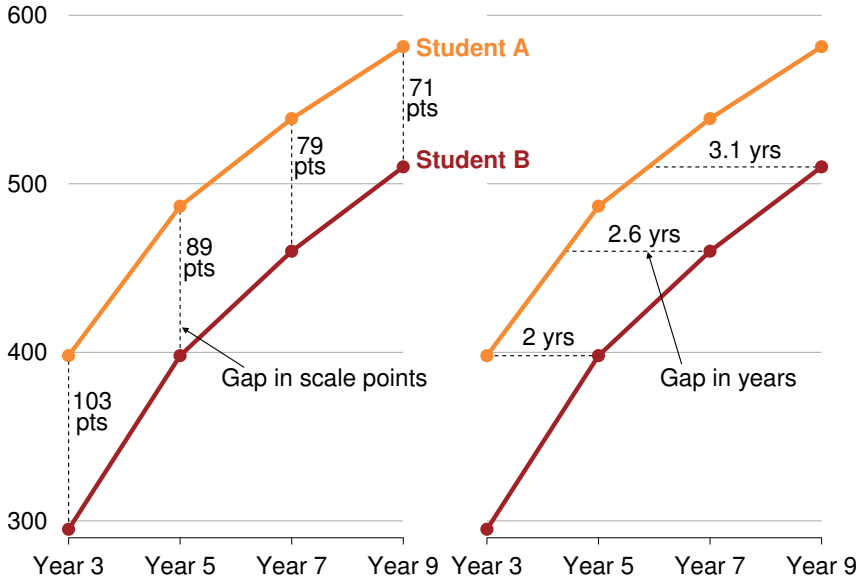
[11] See also Figure 2 and Section 1.5 in the main *Widening gaps* report, Goss et al. (2016).
[12] This type of analysis could also be performed for students that have the same end score.

A.2.2 Looking at progress through the lens of time

An alternative measure of student progress is to define a *year of progress* as the improvement expected from a typical student over a year. This measure would take into account that the typical student makes smaller gains in NAPLAN scale scores as they move further up the NAPLAN scale. That is, the NAPLAN gain score required for the typical student to make two *years of progress* (in terms of the literacy and numeracy skills tested by NAPLAN) between Years 5 and 7 would be smaller than that required between Years 3 and 5.

Years of progress is a measure of student progress relative to their peers, rather than a measure of their absolute *skill level*. This measure gives NAPLAN results new meaning. It can also suggest a very different interpretation of what is happening compared to a 'face value' intrepretation of gain scores or gaps in NAPLAN scale scores.[13] Consider two distinct groups of students: Group A and Group B. The scores displayed on Figure A.5 are those of a representative student within each group (the median student): call these students A and B. Student A scores close to the average for numeracy, while Student B is below average, 103 NAPLAN points behind Student A in Year 3. Looked at in terms of NAPLAN points, as shown on the left chart, the gap between the students has reduced from 103 points in Year 3 to 71 points in Year 9.[14] At face-value, this suggests that Group B are catching up to Group A.

Figure A.5: Measuring progress in years suggests a very different interpretation of NAPLAN results
NAPLAN scale score



Notes: The data in the charts is hypothetical, but the points on both charts are identical.
Source: Grattan analysis.

Yet the chart on the right tells a different story. In Year 5, Student B is performing at the level of Student A in Year 3. But by the time they reach Year 9, Student B's score is roughly half way between Student A's scores in Year 5 and Year 7: Student B is performing at about the level of Student A in Year 6. This suggests that Group B has made about one *less* year of progress than Group A between Years 5 and 9. Looking at progress through the lens of time suggests that Group B are falling further behind, not catching up.

[13] In longitudinal comparisons of two groups of students, changes over time in gaps between the groups are directly related to differences in gain scores.
[14] This does not account for within-group variation, but it suggests the typical student in Group B is catching up to the typical Student in Group A: Student B has a larger gain score between Year 3 and Year 9 than Student A.

A.3 Measuring *Years of Progress*

If we interpret the difference between students A and B according to the chart on the right of Figure A.5, then Student B makes roughly the same progress over four years (between Year 5 and Year 9) as Student A makes in three years (between Year 3 and Year 6). The difference between the students is defined in terms of Student A's rate of learning, but it could just as easily be defined in terms of Student B's rate of learning: "how long will it take Student B to reach the level of Student A?". While the story – that Student A learns comparable skills in less time than Student B – remains the same regardless of which student is defined as the benchmark, the size of the gap between the two in terms of 'years and months' is different. In Year 5, for instance, Student B is performing at Student A's level two years earlier, but Student B will take about three years to reach Student A's current level. We could say that Student A is two years ahead, but we could also say that Student B is three years behind. To compare progress in terms of years and months across different groups requires a common benchmark.

If NAPLAN scores were linked to absolute curriculum standards that define the expected capabilities for each year level, this would provide a common benchmark for measuring progress in terms of time. But given such standards have not been developed, we define a relative benchmark instead.

The results presented in *Widening gaps* use the median or 'typical' student's results as a benchmark for comparing other groups of students. That is, a year of progress is defined according to the gain score expected from the median student at a given level if they were to take the NAPLAN test today and again in one year's time.[15]

NAPLAN scale scores are mapped onto the path of progress of the typical student across their schooling years. We define the schooling year associated with each NAPLAN score as an *equivalent year level*. This type of measure is not new. For instance, the *Programme for International Student Assessment* (PISA) reports a relative grade measure to compare students within each country.[16] Grade equivalent scales have also been used in reporting results of other standardised tests.[17]

It is straightforward to estimate the score corresponding to equivalent year levels 3, 5, 7, and 9; these are the observed median scores for each test-taking year. In Year 5 numeracy in 2014, for instance, the median NAPLAN scale score is approximately 489. A student with a score of 489 in any test-taking year is said to be performing at equivalent year level 5 (using 2014 as a reference year), meaning that their numeracy skill level is the same as a typical Year 5 student.

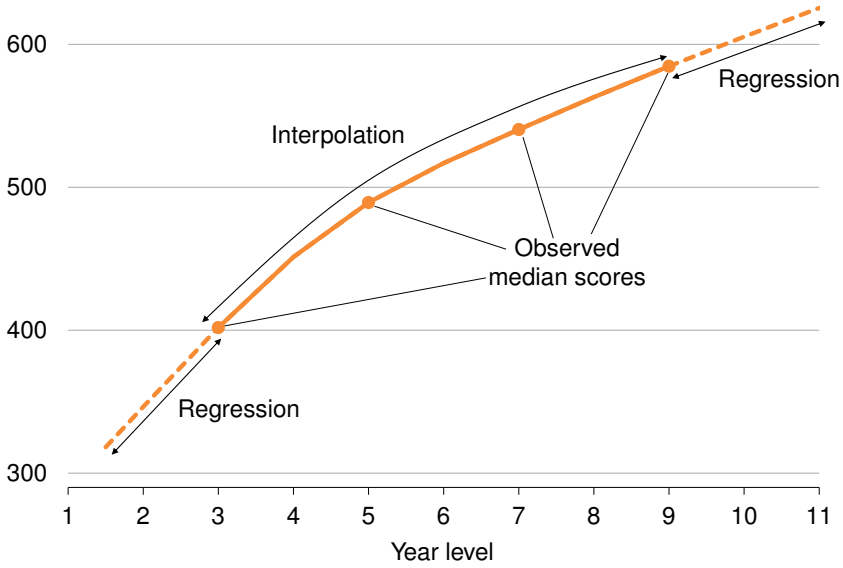
To estimate the median NAPLAN scale score for year levels between Year 3 and Year 9, we fit a curve through the estimated points for Years 3, 5, 7 and 9. This assumes that median student learning follows a smooth trajectory, as opposed to coming in short bursts.[18]

[15] Because NAPLAN is taken every two years, it is only possible to observe gain scores over two-year periods. But it is straightforward to interpolate this for a single year of progress.
[16] OECD (2013).
[17] See, for instance, Renaissance Learning (2015).
[18] This might not be the case for an individual student, but it is a reasonable assumption for the median of a large group of students.

To estimate the median NAPLAN scale score below Year 3 or above Year 9 is more challenging. Without data on Year 2 students, for instance, it is difficult to estimate the skill level of a typical Year 2 student in terms of the NAPLAN scale. But linked data – for instance, Year 3 results in 2012 linked to Year 5 results in 2014 – can be used to estimate the relationship between the Year 3 score and the two-year gain score. Some students will have scored at about the level of a typical Year 4 student on the Year 5 test, meaning they are estimated to be one year behind the typical Year 5 student. We assume that these students were, on average, one year behind the typical Year 3 student when they were in Year 3. Similarly, using Year 7 results linked to Year 9 results, we assume that students who are two years ahead in Year 7 are two years ahead in Year 9, on average.

Using a regression approach, we estimate the median NAPLAN scale score for students who are as much as 18 months behind in Year 3 (which we refer to as equivalent year level 1.5, or Year 1 and 6 months), and as far as 24 months ahead in Year 9 (which we refer to as equivalent year level 11).[19] It is possible to extrapolate the curve further than two years ahead of Year 9, but the results are less robust at such points. Figure A.6 shows conceptually how these approaches are used to construct a curve that maps NAPLAN scale scores to estimated equivalent year levels. The methodology is described in more detail in Appendix C.

Figure A.6: Estimating the equivalent year level benchmark curve involves interpolation and regression
Estimated median NAPLAN scale score, numeracy, Australia



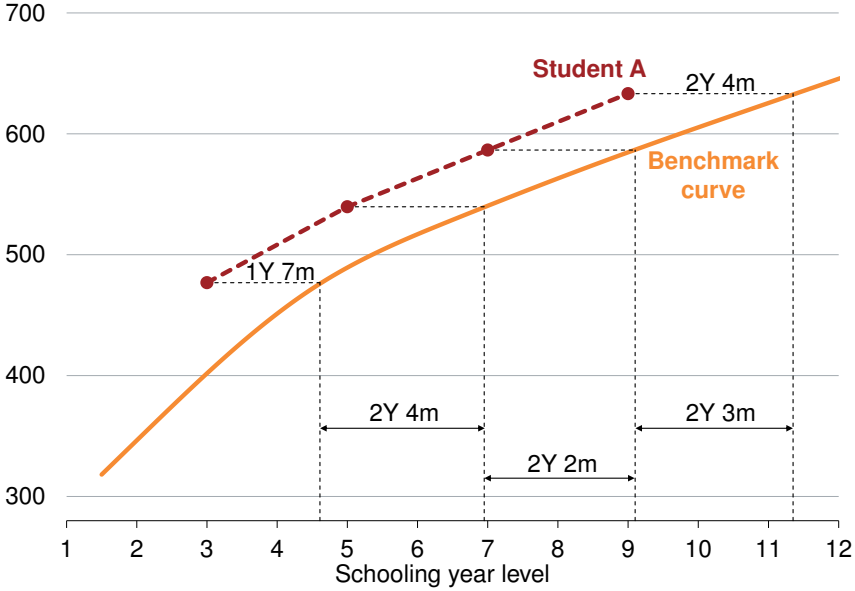
Source: Grattan analysis of ACARA (2014).

[19] It is important to emphasise that this approach is based on data from below-average students in Year 3 and above-average students in Year 9; it is not an extrapolation of the curve constructed between Year 3 and Year 9.

Having constructed the benchmark curve, it is possible to track the equivalent years of progress made by a given student or a group of students. An example of this is shown in Figure A.7 for the median student (Student A) of an above-average student group. In Year 3, this student is about one year and seven months ahead of the benchmark curve in Year 3. By tracking Student A back to the benchmark curve, we can conclude that this group made above-average progress between each NAPLAN test, finishing Year 9 two years and four months ahead of the benchmark. That is, this student made six years and nine months of progress between Year 3 and Year 9.

Figure A.7: Student progress is measured with reference to the benchmark curve

NAPLAN scale score, numeracy, Australia



Source: Grattan analysis of ACARA (2014).

Box A.2: How to interpret equivalent year levels

Equivalent year levels are a meaningful way of comparing the relative progress made by different sub-groups of students. Measuring progress in years also has an intuitive interpretation not available from NAPLAN gain scores.

Yet equivalent year levels should not be over-interpreted. For instance, some Year 5 students are performing at equivalent year level 9 in numeracy – this does not mean these students would necessarily perform comfortably in mathematics at a Year 9 level. In fact, given that these students have not typically been taught the standard mathematics content between Years 6 and 8, we might expect them to struggle with Year 9 content.

A better interpretation is to say that students at equivalent year level 9 have a skill level that is about four years ahead of the typical Year 5 student. That is, the typical Year 5 student is expected to take about four years to reach the skill level of these students at equivalent year level 9. It may be more statistically pure to construct a separate curve for each year level, and interpret all the results relative to that year level (for example, *one year below*, *two years ahead*), but this also adds a layer of complexity to the interpretation of the analysis.

The interpretation of equivalent year levels below 3 requires care. NAPLAN tests are not designed for students below Year 3. While many students in Years 1 and 2 may have comparable reading or numeracy skills to the typical Year 3 student, we do not know how well the typical Year 1 or Year 2 student would perform on a Year 3 NAPLAN test.

The interpretation of equivalent year levels above 9 is even more challenging. For instance, while we would interpret students at equivalent year level 11 in numeracy to be two years ahead of the typical Year 9 student, it is not clear whether the typical Year 9 student will reach this skill level in the next two years. This is because subject choices become more specialised (in many states mathematics is not compulsory in Year 11), and it is possible that the skill level of many students will stagnate after Year 9. It may be more correct to say that the typical Year 9 student would take two years to reach equivalent year level 11 if they continue to study numeracy in a similar way over the next two years.

Some may argue that without data on students below Year 3 and above Year 9, equivalent year levels should not be estimated beyond this range. While there are challenges in estimating and interpreting equivalent year levels outside the Year 3 to Year 9 range, many sub-groups of students score outside this range. Restricting results to this range would severely limit our understanding of student progress. For instance, each comparison would need a specific benchmark, or otherwise we would only be able to compare relative progress between Years 5 and 7, rather than between Years 3 and 9.[a] For policymakers, it is much more useful if student progress can be compared across the majority of schooling years using a single benchmark curve.

[a] The broader policy implications and recommendations of *Widening gaps* would be unchanged even with this much narrower interpretation, because the general patterns of student progress that we find between Years 3 and 9 are consistent with what we find between Years 5 and 7.

B Data sources and issues

B.1 Student-level NAPLAN datasets used in the report

The analysis in *Widening gaps* is based on linked student-level NAPLAN records.[20] There are two major datasets used in the analysis:

- NAPLAN results across all four assessment domains and year levels for all Australian students recorded in 2014, linked with their 2012 results where applicable.[21] This dataset contains test scores for more than one million students for each domain in 2014, and more than 700,000 in 2012.[22]
- NAPLAN results across all four domains recorded between 2009 to 2015 for the cohort of Victorian students who were in Year 3 in 2009.[23] For each domain, more than 55,000 students have a Year 3 test score and a score from at least one other test year. More than 45,000 students have a test score recorded in all of Years 3, 5, 7, and 9 for both reading and numeracy.

Equivalent year levels are estimated using the national dataset to create a national benchmark for student progress. This benchmark is used in analysis of the linked Victorian data, which allows progress of individual students to be tracked from Year 3 to Year 9. In this way, the "years of progress" made by particular groups of Victorian students is relative to the typical Australian student, as opposed to the typical Victorian student.[24]

The data contain a number of student background variables, including gender, parental education and occupation, language background and indigenous status. Some geographic information is available at the school level, including state, and whether the school is located in a metropolitan, regional, or rural area. The Victorian data also include the local government area of the school as well as a measure of school socioeconomic status (SES): the Index of Community Socio-Educational Advantage (ICSEA).[25] The national dataset contains a randomised school-level indicator, but not possible to identify schools themselves.

Two additional datasets tracking different cohorts of students are used to check the robustness of the analysis – the NAPLAN results across all domains and year levels for all Australian students recorded in 2013, linked with their 2011 results, and the NAPLAN results across all domains recorded between 2008 to 2014 for the cohort of Victorian students who were in Year 3 in 2008.[26] Because NAPLAN results vary across cohorts, the analysis was rerun with these data. This confirmed that the key findings of the report – in terms of the scale and direction of learning gaps – were not cohort-specific (see Sections C.4.4 and D.4.3).

[20] Analysis was carried out for reading and numeracy, but not the other domains.
[21] ACARA (2014).
[22] Only students in Years 5, 7, and 9 in 2014 have a linked record in 2012. Linked records are not available for students in the Northern Territory.
[23] VCAA (2015).
[24] This allows the analysis to pick up Victorian-specific effects. It should be noted that, on average, Victorian students score higher than most other states. One explanation for this is that Victorian students are, on average, more likely to come from a high SES background [ACARA (2014)].
[25] To prevent school identification, the Victorian ICSEA data were given to us in bands of 26 points.
[26] ACARA (2013) and VCAA (2014).

B.2 Defining the 'typical' student

The analysis presented in *Widening gaps* focuses on the 'typical' student, either at the population level or within a particular sub-group of students. As noted in the main report and in Appendix A, for the purposes of measuring *Years of Progress*, the typical student in a given year level is defined as the student with the median NAPLAN scale score. Analysis of particular sub-groups of students (such as those grouped by parental education or school ICSEA) is performed according to the typical student within each sub-group – the sub-group median.

An important advantage of using the median over the mean is that it is not directly affected by outliers. For instance, there may be a number of students who do not care about NAPLAN results who leave questions unanswered on the test instead of attempting them, meaning that their NAPLAN scale scores would not be an accurate estimate of their true skill level. These inaccurate results would have a much larger impact on estimates of the mean score and the mean gain score than they would have on the median.[27] NAPLAN scale scores also tend to have a small positive skew (particularly for numeracy), which lifts the mean relative to the median.

B.3 Indicators of parental education

The report analyses how NAPLAN results and progress vary by different levels of parental education, using the Victorian 2009–15 dataset. While there is information on the highest schooling year attained, most parents of school-age children in Victoria have completed Year 12; we therefore focus on educational attainment beyond school. Students can be divided into four groups based on the highest level of post-school parental education:

- at or above Bachelor's degree
- diploma
- certificate I to IV
- no post-school education (Year 12 or below).

Parental education is a strong predictor of household income, and is highly correlated with other socioeconomic factors.[28] For example, in 85 per cent of the households where a parent has a university degree, the highest level of parental occupation is either manager or professional, compared to only 21 per cent of households where neither parent has a degree or diploma.[29]

Figure B.1 on the following page shows that the four categories of parental education include at least 15 per cent of all students. Preliminary analysis suggests that the difference in student attainment and progress between the lowest two categories of parental education is small – as a result, in the main report we group these into a single category: 'below diploma'.

Much of the exploratory analysis in this technical report groups students by parental education.[30]

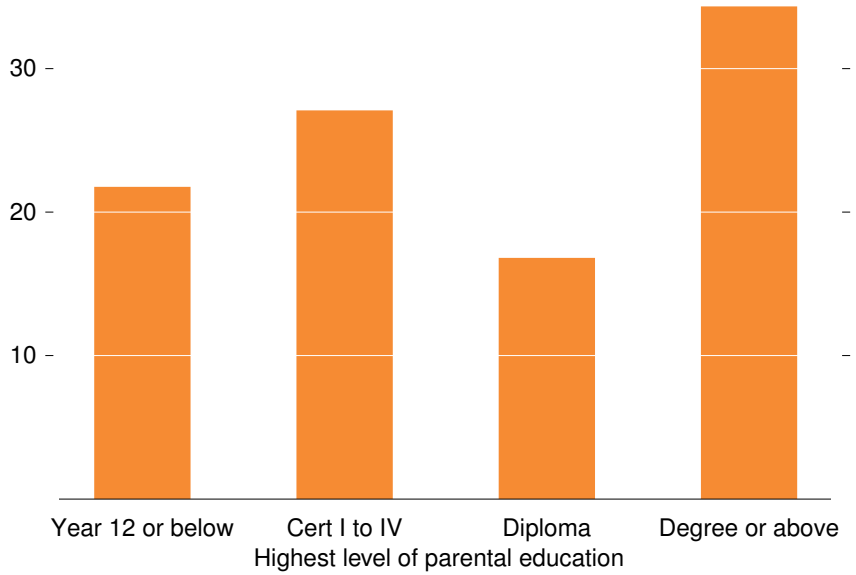
[27] Estimates of the median would only be impacted in this way if a substantial number of students whose true skill level is above the median are recorded below the median as a result of leaving questions unanswered.

[28] See, for instance, OECD (2015).

[29] Grattan analysis of VCAA (2015). Some studies use a composite measure of socioeconomic status, which includes both parental education and occupation, such as Marks (2015) and Houng and Justman (2014). Our analysis highlights the cycle of educational disadvantage, looking at the relationship between parental education and the education outcomes of students.

[30] This is only one of the ways in which we group students to analyse student progress in *Widening gaps*. We explore the robustness of results by parental education for simplicity and consistency.

Figure B.1: Students are well represented in each category of parental education
Percentage of students, Victorian 2009–15 cohort



Source: Grattan analysis of VCAA (2015).

B.4 Using ICSEA as a measure of school socioeconomic status

The report analyses how NAPLAN results and progress vary by the Index of Community Socio-Educational Advantage (ICSEA, which is referred to in the report as 'school advantage') in the Victorian 2009–15 dataset. ICSEA was developed by ACARA so that NAPLAN results could be compared between schools with similar student backgrounds. The index is based on student-level factors such as parental education and employment, indigenous status, and school-level factors such as remoteness and the proportion of indigenous students.[31] The index is constructed as a linear combination of these student- and school-level SES variables.

To determine the weighting applied to each variable, a regression model is estimated: average NAPLAN score (across all domains) against each SES variable. The estimated parameters of this model determine the weightings – essentially this means that the SES variables are weighted according to how strongly they relate to NAPLAN results. This index is then averaged across all students in each school, and scaled nationally so that the ICSEA distribution has a mean of 1000 and a standard deviation of 100. This methodology provides an estimate of ICSEA for each school, which is adjusted each year.[32]

There is a question as to whether this methodology means that we will observe a strong relationship between school ICSEA and NAPLAN results, even if the school SES variables are only weakly related to NAPLAN results. We do not believe this to be the case. While NAPLAN results are used in the construction of ICSEA, they are not used as an input variable – ICSEA is still entirely a linear function of SES variables. That is, the strong relationship observed between ICSEA and NAPLAN results is driven by SES factors, not by the way the index is constructed.

We use the Victorian linked data to analyse the impact of school ICSEA on student progress. We allocate schools into one of three ICSEA groups:[33]

- ICSEA greater than 1090 (approximately the top quartile of schools in Victoria)

[31] Geographic census data are also used in the index calculation.

[32] For more detail, see ACARA (2015a).

[33] Allocation is done for each of 2009, 2011, 2013 and 2015, since schools can change their socio-economic mix and ICSEA is recalculated by ACARA for all schools each year.

- ICSEA greater than 970 but less than 1090 (approximately the middle two quartiles of schools in Victoria)
- ICSEA less than 970 (approximately the bottom quartile of schools in Victoria).[34]

These are referred to as *high advantage*, *medium advantage*, and *low advantage* schools respectively.

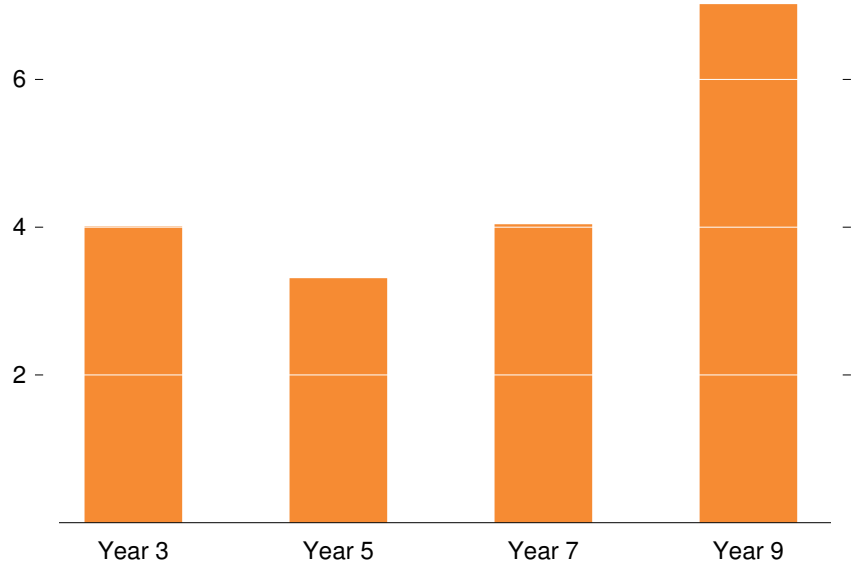
B.5 Missing data

There are two major sources of missing NAPLAN data: non-participation in NAPLAN and results that are not linked for the same student in different years. The non-linkage of results is only an issue for students in the Northern Territory – no linked data are available for Northern Territory in the national dataset.

For any given NAPLAN test, participation rates are high, usually exceeding 90 per cent. The most common reason for non-participation is student absenteeism. This is usually four per cent or less, but rises to seven per cent in Year 9, as shown for numeracy in Figure B.2. A small proportion of students (typically less than two per cent) are given an exemption from taking the NAPLAN test, usually if they have a significant disability or face a major language barrier. Finally, some students are withdrawn from testing by their parent/carer, although this is less than two per cent on almost every test.

Despite a high participation rate on each test, these missing data can potentially reduce the size of the linked samples quite significantly. In the cohort of Victorian students who took the Year 3 test in 2009, only about 72 per cent took all four NAPLAN tests to Year 9 for numeracy and reading. This is because

Figure B.2: Students are more likely to be absent from a NAPLAN test in Year 9
Percentage of students that are absent from NAPLAN numeracy test, Victorian 2009–2015 cohort



Notes: Does not include students who are exempt, withdrawn or miss a test due to leaving Victoria. Results are similar for reading.
Source: Grattan analysis of VCAA (2015).

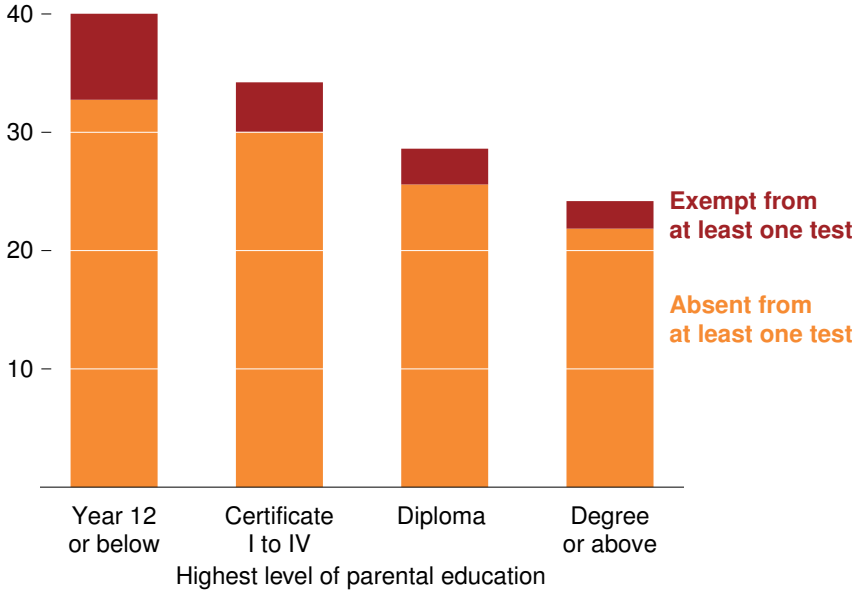
different students missed the test in different years, and also because some students moved out of Victoria before Year 9.[35]

A brief analysis suggests that students are less likely to miss a test due to being absent/withdrawn or an exemption if their parents are better educated. Figure B.3 shows that of the Victorian cohort of students in Year 3 in 2009, 40 per cent of those whose parents have no tertiary education missed at least

[34] These cut points were chosen from the ICSEA bands available to us. It should be noted that the average ICSEA of Victorian schools is about 30 to 40 points higher than the national average.

[35] There are also students that accelerated or repeated a year – these students are included in the analysis, although some have not completed Year 9 by 2015.

Figure B.3: Students from households with higher parental education are less likely to miss one or more NAPLAN tests
Percentage of students that miss a NAPLAN test, Victorian 2009–15 cohort



Notes: Includes all Victorian students in Year 3 in 2009, and all NAPLAN tests taken up to 2015. 'Absent from at least one test' includes those who were withdrawn, and those not in Victoria in one or more test-taking years after Year 3. Students that have been both absent and exempt from tests are categorised as exempt.
Source: Grattan analysis of VCAA (2015).

one test between Year 3 and Year 9, compared to only 25 per cent of students where a parent has a university degree.

Given that students of well-educated parents typically score higher and make higher gains from a given starting score than those whose parents are less well educated, the consequence of ignoring missing data is an upwards bias in estimates of the median score and median gain score.[36]

[36] That is, the estimated median is likely to be above the actual population 50th percentile.

It is also possible that students who miss a test would have made a lower gain score than other students, even after controlling for starting score. With only two years of linked data it would not be possible to test this. But with four years of linked data, as is available with the Victorian 2009 to 2015 cohort, there are students that have missed a test in one or two years, but for whom we observe NAPLAN scale scores in at least two other years. Figure B.4 on the next page shows the estimated median gain score in reading between Years 5 and 7 for students that did not miss a test in any year, and for students that missed a test in Year 3, Year 9 or both. Not only are those that missed a test predicted to make smaller gains, but the gap is larger for students whose parents do not have a degree or diploma.

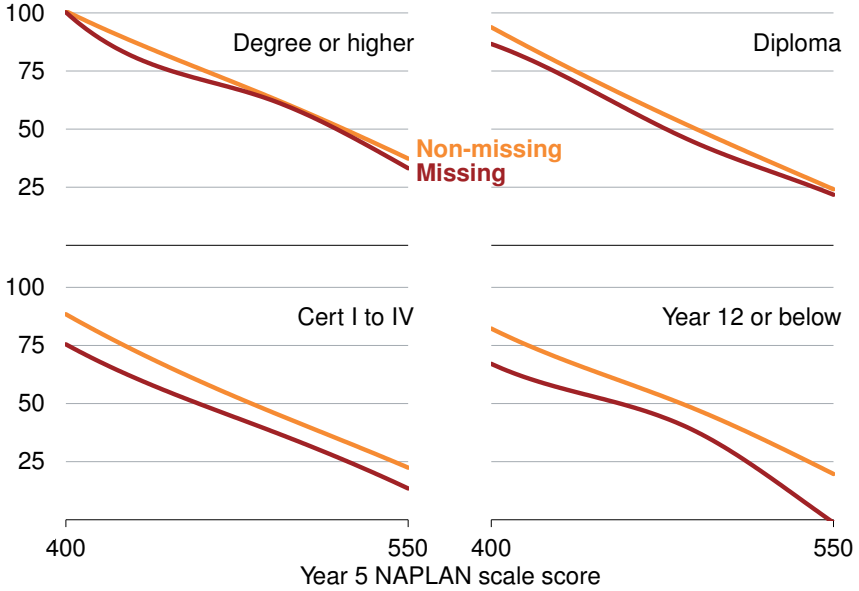
This means that estimates of median progress for particular sub-groups are likely to be upwards biased if missing data are ignored. But the bias is likely to be much larger for lower levels of parental education. In turn, this means the gap in student progress calculated between students with high and low parental education is likely to be underestimated rather than overestimated.[37]

Our analysis of NAPLAN gain scores does not impute missing results. Students who are given an exemption from one or more tests are excluded from the analysis.[38] When estimating progress for Victorian students, we aim to minimise bias –

[37] The report shows a very consistent pattern of students from well-educated households out-performing those from lower-educated households in Year 3, and this gap growing over time. A similar pattern is found between high and low advantaged schools. These are key findings of the report. If missing data could be adequately taken into account, it is likely that these gaps would be estimated to be even larger.

[38] For the purposes of reporting, ACARA assume exempt students are performing below the national minimum standard. Imputing NAPLAN scale scores for these students would change the sample median, but with so few students exempt it is unlikely the results would change significantly.

Figure B.4: Missing data have more of an impact on gain scores for students from less-educated households
Median NAPLAN gain score by highest level of parental education, reading, Year 5 to Year 7, Victorian 2009–15 cohort



Notes: 'Missing' includes all students that were absent/withdrawn from either the Year 3 or Year 9 reading test, but does not include exempt students. 'Non-missing' includes all students that did not miss a single NAPLAN test. A similar pattern exists for numeracy, for other year levels, and for school advantage.
Source: Grattan analysis of VCAA (2015).

rather than excluding all students that miss a test, we include all students that undertook the Year 3 test and at least one other test. This approach is outlined in more detail in Section D.2.1.

B.6 Measurement error and bias

B.6.1 Measurement error at the student level

The NAPLAN scale score that a student receives for a particular test is known as a 'weighted likelihood estimate' (WLE).[39] Two students that answer the same number of correct answers on the same test receive the same WLE.

The score that a student receives on the NAPLAN test provides an estimate of their true skill level in a particular domain, but this is subject to substantial measurement error. The accuracy of the estimate increases with the number of questions asked.[40] Two scores are needed to estimate progress over time, and each is subject to measurement error. It is therefore difficult to accurately estimate the progress of an individual student using NAPLAN.

NAPLAN results are more accurate for estimating the progress of a sizeable group of students, as measurement error is reduced when results are aggregated across students. But simply aggregating does not solve all of the potential measurement error issues. This section outlines these issues in detail and explains the approach we have taken to mitigate them.[41]

[39] These are also referred to as 'Warm's Estimates'; see Warm (1989).
[40] On the Year 3 numeracy test in 2009, for instance, there are 35 questions, and NAPLAN scale scores are estimated with a standard error between 24 and 35 for the vast majority of students. On the Year 9 numeracy test in 2015, there are 64 questions, and the standard error of NAPLAN scale scores is between 17 and 30 for nearly all students. Extreme scores (nearly all questions correct/incorrect) are estimated with much higher standard errors [ACARA (2015d)].
[41] There may also be measurement error issues in other variables – for instance, parental education may change over the course of a child's schooling years, but this is not recorded. Our analysis assumes that the recording of background variables is accurate.

B.6.2 Using NAPLAN scale scores (WLEs) may result in imprecise estimates of progress

Skill level is continuous, but NAPLAN scale scores are discrete

NAPLAN scale scores provide an estimate of student skill level, a continuous latent variable. But because there are a finite number of questions on each NAPLAN test, the estimates of student skill level (NAPLAN scale scores) have a discrete distribution.

On the Year 3 numeracy test, for example, there are only 35 questions, meaning that there are only 35 possible NAPLAN scale scores a student can receive. The cohort of students that takes the test in 2014 would receive a different set of scores to the cohort taking the test in 2015, even where there is no significant difference between the two cohorts.[42] Ignoring the discrete nature of the estimates could overstate the difference between two cohorts because of 'edge effects', especially when comparing performance in terms of percentiles, such as the progress or achievement of the median student.

Regression to the mean

In the context of comparing student progress over two or more NAPLAN tests, *regression to the mean* suggests that an extreme NAPLAN score in one year (either extremely low or high) is likely to be followed by a less extreme score on the following test (two years later). This is not because students at the extremes are making significantly high or low progress, but because the original test score is exaggerated by measurement error. This may lead to learning progress being significantly overstated by gain scores for students who start with a very low score, and understated for students who start with a very high score.[43]

Wu (2005) notes that the average of the WLEs provides an unbiased estimate of the population mean skill level, but the sample variance overstates the population variance. This bias disappears as the number of test questions increases. For students who score close to the mean, the bias in the WLE as an estimate of their skill level will be small. But for extreme percentiles, the bias can be large.[44]

It is important to note that an extreme score for a particular sub-group might not be an extreme score for another sub-group. For example, the NAPLAN scale score equal to the 95th percentile in Year 7 numeracy for those whose parents have no post-school qualifications is only at the 82nd percentile for those who have a parent with a university degree. This means that the regression to the mean between the Year 7 and Year 9 test is likely to be stronger for a high achieving student whose parents have no post-school qualifications than it is for a high achieving student with a university-educated parent.[45]

[42] A histogram comparing two cohorts would show a similar overall distribution, but the estimated points on the NAPLAN scale would be different. It is therefore important to take care when interpreting results across students from different cohorts.

[43] The data show a systematic pattern of high gain scores for low prior scores and low gain scores for high prior scores; see, for example, Figure A.2 on page 8 and Figure B.4 on page 21. But if this were entirely due to regression to the mean, we would expect the path of progress for the median student from Year 3 to Year 9 to be approximately linear – this is clearly not the case.

[44] A way to think about this is that the effective number of questions declines as student skill level moves further from the level at which the test is set. For example, a student at the 90th percentile will find most questions too easy, while a student at the 10th percentile will find most questions too difficult. Only a few questions will be set at an appropriate level for such students. The move to NAPLAN online will allow better targeting of questions, reducing the measurement error at the extremes.

[45] Just because a student does not have a university-educated parent, this does not mean that a high NAPLAN scale score is overstating their true skill level. But when we compare two students with the same high score, one

B.6.3 Approaches to mitigate the impact of measurement error and bias

Simulation approach

All WLEs (NAPLAN scale scores) are point estimates and are associated with a standard error. Warm (1989) shows that these estimates are asymptotically normally distributed. Using this property, we approximate the distribution of student skill level, θ, given these estimates:

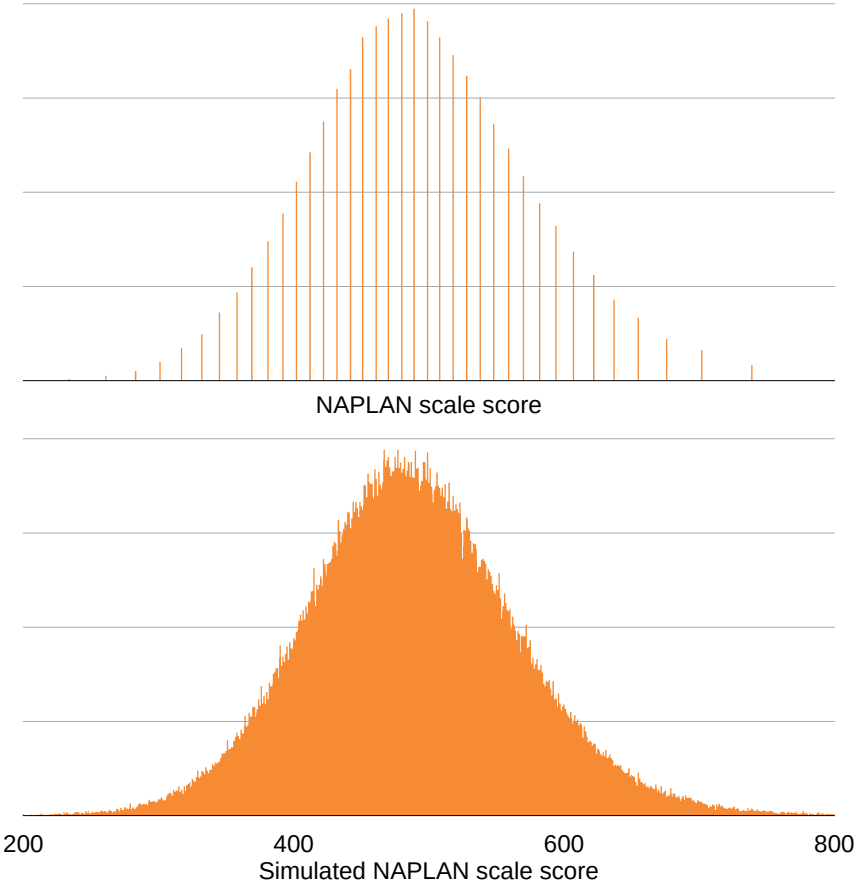
$$\theta_n \overset{a}{\sim} \mathcal{N}\left(\widehat{\mu}_n, \widehat{\sigma}_n^2\right) \tag{B.1}$$

where n is the number of questions correctly answered, $\widehat{\mu}_n$ is the corresponding WLE, and $\widehat{\sigma}_n^2$ is the variance of the WLE.

For each student, we simulate a NAPLAN scale score (skill level) as a random draw from this distribution.[46] This creates a sample that has the properties of a continuous distribution, allowing for more accurate estimates of percentiles.

Figure B.5 compares a histogram of discrete NAPLAN scale scores to a histogram of simulated NAPLAN scale scores. While this approach does not remove measurement error at the individual student level, it takes into account that measurement error varies across students with different scores.

Figure B.5: The simulation approach solves the issue of discrete NAPLAN scale scores
Histogram of Year 5 NAPLAN scale score, numeracy, Australia



Notes: Frequency is not shown on Y-axes, but scaled so that both charts can be compared. Bin width = 0.5.
Source: Grattan analysis of ACARA (2014).

with a university-educated parent and one without, the one without is more likely to have had an unusually good test day (i.e. scoring above their true skill level) than the student with a university-educated parent.

[46] This is performed for each year in the Victorian cohort and each year in the national dataset, using the standard errors reported by ACARA (2015d).

Use of sub-groups with large samples

Simulating NAPLAN scores does not remove measurement error at the individual student level. In fact, it increases the standard error associated with an individual student estimate and gain score.[47] We keep this measurement error to a minimum by aggregating students into sub-groups that have large samples, and calculating our results based on multiple random draws.[48]

Avoiding extreme percentiles

There is no straightforward way to estimate the magnitude of the bias in the WLEs for different percentiles. But it is well known that the magnitude of the bias due to regression to the mean is largest for extreme percentiles, and that the bias is small for percentiles close to the median. The impact of regression to the mean is also larger when the correlation between two measurements (such as test scores) is weak. In our sample, the correlation between NAPLAN test scores across two test-taking years for a given domain is between 0.75 and 0.8 – this strong correlation suggests regression to the mean will have only a small impact for most percentiles.

Nonetheless, our analysis aims to avoid estimating NAPLAN scale scores and gain scores for students at extreme percentiles, and most analysis is focused around the median student. We use a rule of thumb to minimise bias due to regression to the mean – no analysis is based on the estimated NAPLAN scale score or gain score of students below the 10th percentile or above the 90th percentile.[49]

In constructing the benchmark curve to estimate equivalent year levels (outlined in Appendix C on page 26), it is necessary to estimate the median gain score of below-average students from Years 3 to 5, and above-average students from Years 7 to 9. It is possible to estimate the NAPLAN scale score for a student as low as 18 months behind Year 3 level, and as high as three years ahead of Year 9 level without using extreme percentiles.

For the analysis of progress using Victorian data, we track low, medium, and high achieving students based on their percentile at Year 3 – the 20th, 50th, and 80th of the Victorian population. But these percentiles can be more extreme when analysing sub-groups. In Year 3 numeracy, for example, the 20th percentile across the population is equal to the 12th percentile for students who have a parent with a university degree, and the 80th percentile at the population level is the 87th percentile when the highest level of parental education is below a diploma. Table B.1 shows the within-group percentiles for the 20th and 80th percentiles in Year 3 at the population level. Using these percentiles at the population level ensures that we do not go below the 10th or exceed the 90th percentile for any parental education sub-group.[50]

The gaps in progress between high and low parental education levels may still be overstated due to regression to the mean, particular when comparing from the 20th or the 80th percentile in Year 3. This is explored more in Section D.4.2.

[47] This approach would be inappropriate for reporting individual student results.

[48] Sub-groups analysed typically have between 7000 and 25,000 students. The standard error due to measurement in a sub-group is proportional to $\sqrt{n}$, the square root of the sub-group sample size. For a sub-group with 10,000 people, the standard error will be 100 times smaller than it will be for an individual student.

[49] These extreme percentiles are avoided both for the overall population, and for particular sub-groups.

[50] This also holds for school advantage.

Table B.1: Using the 20th and 80th percentiles at the population level avoids extreme percentiles within sub-groups
Within-group percentile in Year 3 numeracy by parental education, Victorian 2009–15 cohort

Sub-group	Percentile	
Population	*20*	*80*
Degree or above	11.9	70.8
Diploma	19.7	81.5
Below diploma	26.3	86.7

Notes: 'Extreme percentiles' defined as below 10th or above 90th within a sub-group.
Source: Grattan analysis of VCAA (2015).

Reporting of results and standard errors

To simplify the presentation of our findings, the report does not show standard errors on point estimates of NAPLAN scale scores or equivalent year levels. But confidence bounds are estimated to ensure the significance of reported results. We calculate 99 per cent confidence intervals using a bootstrap approach with 200 replications, each with a different set of random draws.[51] Separate bootstrap simulations are run for estimation of the benchmark curve with the national dataset and for estimation of student progress using the Victorian dataset.

We estimate a confidence interval for the benchmark equivalent year level curve, as well as confidence intervals for the analysis of progress using the Victorian cohort. For results that are reported in terms of equivalent year levels or years of progress, these confidence intervals are calculated using both bootstrap simulations.[52]

The confidence intervals are used to validate the significance of our findings – we do not draw conclusions from any results that are not statistically significant (at the 1 per cent level).

Plausible values

The best approach to reduce the impact of measurement error is to use *plausible values*. Like the simulation approach outlined above, this approach would simulate a NAPLAN scale score from a continuous distribution for each student, including imputing values for missing data. But plausible values are simulated from a distribution that takes into account student and school background factors.[53] NAPLAN reports produced by ACARA are based on analysis using plausible values.[54]

When simulated correctly, plausible values are able to produce unbiased estimates of percentiles and gain scores for each sub-group.[55] Plausible values were available for the 2014 test year in the national dataset, but not for the 2012 results or the Victorian 2009–15 cohort. This means we did not have the data to use plausible values to analyse progress.[56]

We do, however, utilise the 2014 plausible values (generated by ACARA) for estimating the population distribution of results for each year level. These estimates therefore take missing data and measurement error into account.

[51] The lower bound of each confidence interval is estimated as the average of the two smallest bootstrap point estimates, while the upper bound is estimated as the average of the two largest bootstrap point estimates.

[52] Each replication from one simulation is linked to a replication from the other. This approach takes into account the measurement error in the Victorian cohort, as well as the measurement error in the estimation of equivalent year levels.

[53] In theory these could also take into account NAPLAN scores in other year levels.

[54] ACARA (2015e), p. 22.

[55] Wu (2005).

[56] In any case, the 2014 plausible values are, to the best of our knowledge, generated independently of prior test scores. Analysing student progress would ideally be done using plausible values simulated from a distribution that takes both prior and subsequent test scores into account.

C Methodology for mapping NAPLAN scale scores to equivalent year levels

C.1 Introduction

The NAPLAN scale is designed to be independent of year level – a student should receive the same score on average regardless of whether they take a test normally administered to Year 3, Year 5, Year 7 or Year 9 students.[57] This property makes it possible to compare students in different test-taking year levels. For example, a Year 5 student is predicted to be reading above the typical Year 7 level if they score higher than the typical Year 7 student in NAPLAN reading. But because NAPLAN tests are only administered to students in four different year levels, it is not possible to compare students to those outside these year levels without further assumptions.

Widening gaps presents a new framework from which to interpret NAPLAN results. NAPLAN scale scores are mapped onto a new measure, *equivalent year levels*. The NAPLAN scale score corresponding to the equivalent year level 4, for example, is the median score expected from students if they took an age-appropriate NAPLAN test when they were in Year 4.[58]

This appendix outlines the theoretical framework for mapping NAPLAN scale scores onto equivalent year levels and the methodology and assumptions used to estimate this relationship.

C.2 Theoretical framework for mapping

Let X_j ($X_j \in \mathbb{R}$) be a random variable denoting student skill level (as estimated by NAPLAN scale scores) in domain j ($j =$ reading, numeracy), and Y be a variable denoting schooling year level, continuous over the range of schooling years, (y_{min}, y_{max}).[59]

We assume that median student skill level increases monotonically as students progress through school. We define a function $f_j(Y)$ as the median of X_j conditional on Y:

$$\begin{aligned} f_j(Y) &= Q_{50}\left[X_j \mid Y\right] \\ y_1 < y_2 &\implies f_j(y_1) < f_j(y_2) \\ f_j(Y) &\in [f_j(y_{\min}), f_j(y_{\max})] \end{aligned} \tag{C.1}$$

That is, $f_j(Y)$ is the median NAPLAN scale score in domain j of students taking a NAPLAN test in year level Y. For every schooling level there is a corresponding median NAPLAN scale score (for each domain). We also assume that $f_j(Y)$ is continuous and monotonically increasing – at the population level, median student skill level increases steadily over time.[60]

[57] A student's NAPLAN scale score will generally be a more precise estimate of their true skill level when they are administered an age-appropriate test. Giving a typical Year 3 student a test meant for Year 9 students is likely to produce a NAPLAN scale score with a large standard error.

[58] To be precise, in May of the year they were in Year 4, as this is when the NAPLAN test is taken.

[59] Lower case letters are used to denote realisations of these random variables. This report's analysis focuses on reading and numeracy only, but it would be possible to apply the same analysis to the other assessment domains.

[60] For example, if NAPLAN tests were taken every month, we would expect the median score to improve with every test. This may not hold for individual students, but should hold at the population level.

Following this, we propose that a given NAPLAN scale score corresponds to a median schooling year – the point in time in the median student's path of progress (in terms of year level and months) at which their skill level is equal to that score. We define this schooling year as an *equivalent year level*, denoted as Y^*:

$$Y^* = f_j^{-1}(X_j) \tag{C.2}$$

All NAPLAN scale scores in the range $[f_j(y_{min}), f_j(y_{max})]$ therefore correspond to an *equivalent year level*.

C.3 Estimating equivalent year levels

This methodology aims to estimate $f_j(Y)$ for reading and numeracy for a range of different year levels, $Y = 1, 2, ..., 12$, then interpolate over these points to construct a smooth curve. If the NAPLAN tests were administered to students in every year level from Year 1 to Year 12, we could estimate $f_j(Y)$ as the sample median from each of these year levels.[61] But with the tests only administered in four year levels, we must make further assumptions to estimate $f_j(Y)$.

The report estimates $f_j(Y)$ (the median NAPLAN scale scores corresponding to a given year level) using the simulated NAPLAN results (see Section B.6.3) of all Australian students in 2014 linked to their 2012 simulated results (where applicable). It is possible to apply this methodology to NAPLAN results in other years, provided linked data are available.

[61] This is a useful way of thinking about what equivalent year levels are trying to measure. But it is important to note that the interpretation of equivalent year levels 11 and 12 estimated with the available data could be very different to those estimated with data on Year 11 and Year 12 students, as explained in Box A.2 on page 15.

Step 1: Estimate the median NAPLAN scale scores at year levels 3, 5, 7, and 9

These are estimated as the sample median scores in those year levels:

$$\begin{aligned} \widehat{f}_j(3) &= \tilde{x}_{j,3} \\ \widehat{f}_j(5) &= \tilde{x}_{j,5} \\ \widehat{f}_j(7) &= \tilde{x}_{j,7} \\ \widehat{f}_j(9) &= \tilde{x}_{j,9} \end{aligned} \tag{C.3}$$

where $\tilde{x}_{j,y}$ is the sample median NAPLAN scale score in year level y.[62]

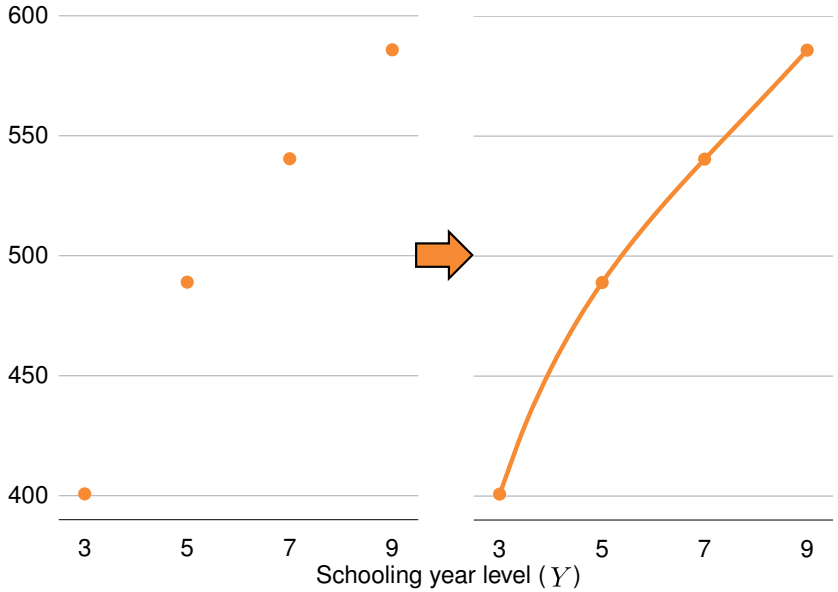
Step 2: Interpolate between Year 3 and Year 9

Using a third-order polynomial, fit a smooth curve through the four data points, $([Y, \widehat{f}_j(Y)], Y = 3, 5, 7, 9)$, to estimate $f_j(Y)$ between Year 3 and Year 9, as shown in Figure C.1.

[62] For Years 3, 5, and 7, we estimated the corresponding NAPLAN scale score, $\widehat{f}_j(Y)$, as the average of the medians in 2012 and 2014.

Figure C.1: A third-order polynomial is used to interpolate between Year 3 and Year 9

Estimated median NAPLAN scale score, $\widehat{f}_j(Y)$, numeracy, Australia



Source: Grattan analysis of ACARA (2014).

Step 3: Estimate the median gain score for Years 3 to 5 and Years 7 to 9 conditional on prior score

To estimate $f_j(Y)$ above Year 9 and below Year 3, we denote a function, $g_{j,Y}\left(X_{j,Y-2}\right)$, equal to the median gain score conditional on year level and a student's NAPLAN scale score from two years earlier:

$$g_{j,Y}\left(X_{j,Y-2}\right) = Q_{50}\left[X_{j,Y} - X_{j,Y-2} | Y, X_{j,Y-2}\right] \qquad \text{(C.4)}$$

where $X_{j,Y}$ denotes NAPLAN scale score in domain j in school year Y. For students that scored $x_{j,3}$ in Year 3 reading, for example, $g_{j,5}(x_{j,3})$ is the median gain score these students will make to Year 5.[63]

From eqs. (C.1) and (C.4), it follows that:

$$g_{j,Y}\left[f_j(Y-2)\right] = f_j(Y) - f_j(Y-2) \qquad \text{(C.5)}$$

That is, the difference between the median scores two years apart is equal to the median gain made from the same starting score.

To estimate $g_{j,Y}$ for $Y = 5$ and $Y = 9$ first requires parameterising the functions. We allow for non-linearity in $g_{j,Y}$ by using restricted cubic regression splines, meaning that $g_{j,Y}$ can be written as a linear function:

$$\begin{aligned} g_{j,Y}(X_{j,Y-2}) = \beta_0 + \beta_1 X_{j,Y-2} + \beta_2 S_2(X_{j,Y-2}) \\ + \beta_3 S_3(X_{j,Y-2}) + \beta_4 S_4(X_{j,Y-2}) \end{aligned} \qquad \text{(C.6)}$$

where S_2, S_3 and S_4 are functions that create spline variables.[64] Alternatively, this function could be specified with quadratic or higher order polynomial terms.

Given $g_{j,Y}$ represents a conditional median gain score, eq. (C.6) can be thought of as a quantile regression model at the median. This can be estimated using least absolute deviations.[65]

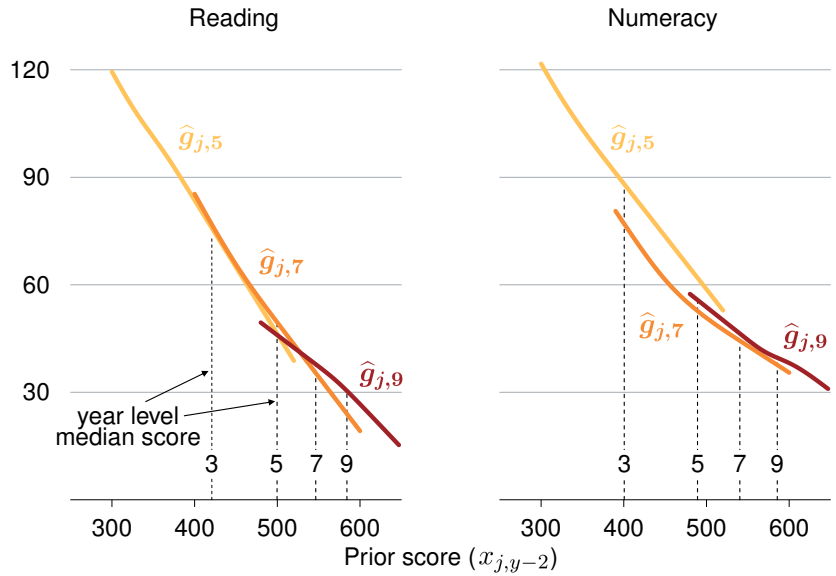
Figure C.2 plots the estimated functions, $\widehat{g}_{j,y}\left(x_{j,y-2}\right)$, for $y = 5, 7$ and 9 for both reading and numeracy. Predicted median NAPLAN gain scores are much higher for lower prior scores, but year level does not have a large effect on gain scores once prior scores are controlled for. For instance, when evaluated at the NAPLAN

[63] The function $g_{j,Y}$ can only be empirically estimated for $Y = 5, 7$ and 9, corresponding to gain scores from Years 3 to 5, Years 5 to 7, and Years 7 to 9 respectively.

[64] More spline variables can be included, if desired.

[65] It is only necessary to estimate $g_{j,5}$ for $x_{j,3} \leq \widehat{f}_j(3)$ and $g_{j,9}$ for $x_{j,7} \geq \widehat{f}_j(7)$.

Figure C.2: The estimated median gain score is strongly related to prior score, but only weakly related to year level
Two-year median NAPLAN gain score, $\widehat{g}_{j,y}(x_{j,y-2})$, Australia



Source: Grattan analysis of ACARA (2014).

score for equivalent year level 3, $\widehat{f}_j(3)$, the functions $\widehat{g}_{j,5}$ and $\widehat{g}_{j,7}$ are extremely close for reading, and similar for numeracy. Similarly, when evaluated at equivalent year level 7, $\widehat{f}_j(7)$, the functions $\widehat{g}_{j,9}$ and $\widehat{g}_{j,7}$ are very close for both reading and numeracy. That is, expected NAPLAN gain from a given starting point is similar for students that are two year levels apart.

Setting $Y = 10$ and re-arranging eq. (C.5) gives:

$$f_j(10) = f_j(8) + g_{j,10}[f_j(8)] \tag{C.7}$$

The point $f_j(8)$ was estimated in Step 2, but it is not possible to estimate $g_{j,10}$ without NAPLAN data for Year 10 students (linked to Year 8 results). But given that year level has little effect on gain scores once prior scores are controlled for, we can assume:

$$g_{j,10}[f_j(8)] \approx g_{j,9}[f_j(8)] \tag{C.8}$$

That is, a student in Year 8 performing at the median Year 8 level will make a similar gain over two years as a Year 7 student performing at the median Year 8 level.

It is necessary to make a stronger assumption to estimate $f_j(11)$:

$$g_{j,11}[f_j(9)] \approx g_{j,9}[f_j(9)] \tag{C.9}$$

That is, we assume a student in Year 9 performing at the median Year 9 level will make a similar gain over two years as a Year 7 student performing at the median Year 9 level.

Similarly, we can use our estimate of $g_{j,5}$ as a proxy for $g_{j,4}$ by assuming:

$$g_{j,4}[f_j(2)] \approx g_{j,5}[f_j(2)] \tag{C.10}$$

That is, a Year 2 student performing at the median Year 2 level is assumed to make a similar gain over two years as a Year 3 student performing at the median Year 2 level.

Step 4: Estimate the median NAPLAN scale scores for year levels 10 and 11

Using the assumptions made in eq. (C.8) and eq. (C.9), $f_j(10)$ and $f_j(11)$ are estimated using the following:

$$\begin{aligned}\widehat{f}_j(10) &= \widehat{f}_j(8) + \widehat{g}_{j,9}\left[\widehat{f}_j(8)\right] \\ \widehat{f}_j(11) &= \widehat{f}_j(9) + \widehat{g}_{j,9}\left[\widehat{f}_j(9)\right]\end{aligned} \tag{C.11}$$

where, for example, $\widehat{f}_j(8)$ is the estimated median NAPLAN scale score for Year 8 students, calculated in Step 2, and $\widehat{g}_{j,9}$ is the estimated median NAPLAN gain score function from Year 7 to Year 9, calculated in Step 3.

Step 5: Estimate the median NAPLAN scale scores for year levels 1.5, 2, and 2.5

Using the assumption made in eq. (C.10) and its extensions, $f_j(1.5), f_j(2)$ and $f_j(2.5)$ are estimated by solving the following equations for $\widehat{f}_j(Y)$:

$$\begin{aligned} \widehat{f}_j(1.5) &= \widehat{f}_j(3.5) - \widehat{g}_{j,5}\left[\widehat{f}_j(1.5)\right] \\ \widehat{f}_j(2) &= \widehat{f}_j(4) - \widehat{g}_{j,5}\left[\widehat{f}_j(2)\right] \\ \widehat{f}_j(2.5) &= \widehat{f}_j(4.5) - \widehat{g}_{j,5}\left[\widehat{f}_j(2.5)\right] \end{aligned} \tag{C.12}$$

where, for example, $\widehat{f}_j(3.5)$ is the estimated median NAPLAN scale score for Year 3 students, six months after the NAPLAN test (November), and $\widehat{g}_{j,5}$ is the estimated median gain score function from Year 3 to Year 5, calculated in Step 3. These points are estimated closer together because $f_j(Y)$ has a larger gradient for lower values of Y.

Step 6: Interpolate over estimated points

Using a range of estimated points for $\left[Y, \widehat{f}_j(Y)\right]$ (for example, use Y = 1.5, 2, 2.5, 3, 4, 5, 6, 7, 8, 9, 10, 11), construct a smooth curve for $\widehat{f}_j(Y)$ using interpolation.[66] Using linear extrapolation, this curve is extended so that $y_{min} = 1$ and $y_{max} = 13$ (Year 13 is reported as 'above Year 12'), although our analysis avoids these extremes as much as possible given the estimates are less robust and standard errors are high.[67]

We now have a curve that estimates the median NAPLAN scale score for each schooling year level: $\widehat{f}_j(Y)$. The inverse of this curve is used to estimate the equivalent year level, Y^*, corresponding to any given NAPLAN scale score, X_j:

$$\widehat{Y}^* = \widehat{f}_j^{-1}(X_j) \tag{C.13}$$

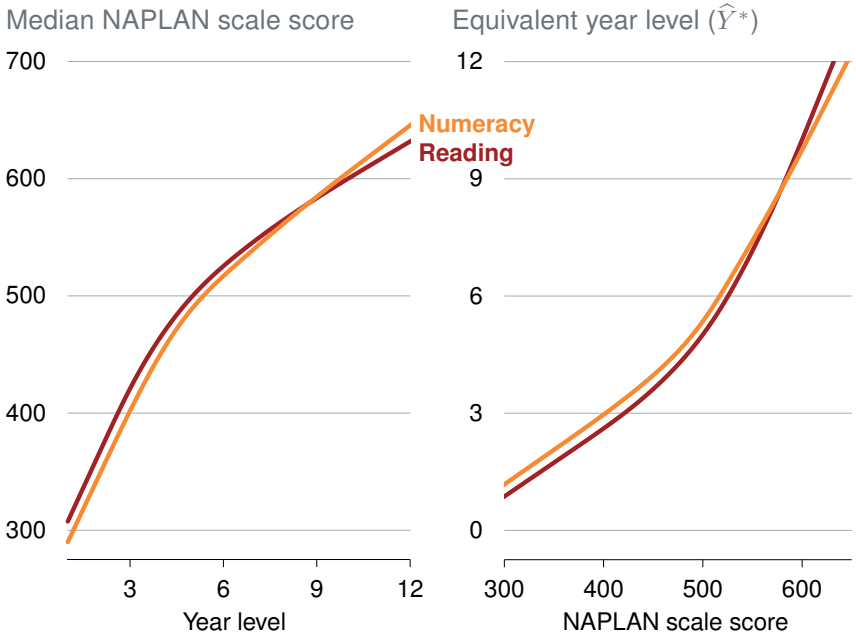
Figure C.3 shows this curve for reading and numeracy, both in terms of $\widehat{f}_j(Y)$ and in terms of its inverse, $\widehat{f}_j^{-1}(X_j)$. As the chart on the right shows, every NAPLAN score (within the range of the curve) can be mapped to an equivalent year level. A score of 500 in numeracy, for instance, corresponds to an equivalent year level of 5 years and 4 months – a student at this level can be interpreted as performing four months ahead of the typical (median) Year 5 student at the time of the Year 5 NAPLAN test.[68]

[66] Our methodology fits a curve using a regression with restricted cubic splines – some of the points already estimated for $f_j(Y)$ shift slightly as a result.

[67] See Box A.2 on page 15 for a discussion about the interpretation of equivalent year levels estimated outside the range of Year 3 to Year 9. Given the estimated curve, $\widehat{f}_j(Y)$ is approximately concave between Year 1.5 and Year 11, we would expect concavity to hold if the curve is extended to Year 1 and Year 13. As such, linear extrapolation is unlikely to underestimate the median scale score for Year 1, Year 12, and Year 13 – this is conservative for estimating the gaps in progress between different groups.

[68] Given that NAPLAN is administered in May of each year, another interpretation is to say that this student is performing at the level we would expect of the typical Year 5 student in September.

Figure C.3: All NAPLAN scale scores in a given range correspond to an equivalent year level



Notes: Left chart shows estimated function $\widehat{f}_j(Y)$, while right chart shows its inverse, $\widehat{f}_j^{-1}(X_j)$. The left chart can be interpreted as the estimated median NAPLAN scale score for a given year level, whereas the right chart can be interpreted as the estimated equivalent year level for a given NAPLAN scale score.
Source: Grattan analysis of ACARA (2014).

These curves can be used to compare different cohorts or sub-groups of students in terms of differences in their achievement, and to track student progress relative to the median student. Years of progress is simply calculated as the difference in equivalent year levels between two points in time. If, for example, a student makes 2 years and 6 months of progress over a two-year period, they have made the same amount of progress as the typical (median) student is expected to make over 2 years and 6 months, starting from the same point.

C.4 Robustness of equivalent year level estimates

There are a number of questions that may arise in relation to the methodology used to estimate equivalent year levels. For instance:

- what is the standard error at different points along the equivalent year level curve?
- how accurate are estimates beyond Year 3 and Year 9?
- how do the estimates change with different assumptions?
- are the results robust to the cohort used?

It is worth investigating each of these questions in detail to ensure that the methodology and the results are robust.

C.4.1 Standard errors around point estimates

There are two sources of error that the standard error accounts for: test measurement error for individuals, and the error associated with a finite sample. But the equivalent year level curve is calculated from a very large sample, meaning that the standard error around estimates of the median is naturally small.[69]

In reporting, we prefer using confidence intervals to standard errors, since equivalent year levels are asymmetrically distributed around NAPLAN scale scores. We calculate a 99 per cent confidence interval at each point along the curve, $\widehat{f}_j(Y)$, between $Y = 1$ and $Y = 13$. This is based on a bootstrap simulation with 200 replications.[70]

Between Year 3 and Year 9, equivalent year levels are estimated with a very narrow confidence interval. As the curve is flatter in Year 9 than it is in Year 3, the confidence interval around Year 9 is wider. The width of the confidence interval naturally increases below Year 3 or above Year 9. For a score of just over 300 in reading (close to equivalent year level 1), the 99 per cent confidence interval around the equivalent year level estimate is about seven months of learning, while for a score of 650 (close to equivalent year level 13), the 99 per cent onfidence interval is eight months.[71] But for scores between 400 and 600, the 99 per cent confidence interval does not exceed two months of learning. These intervals are displayed in Figure C.4 and table C.1.

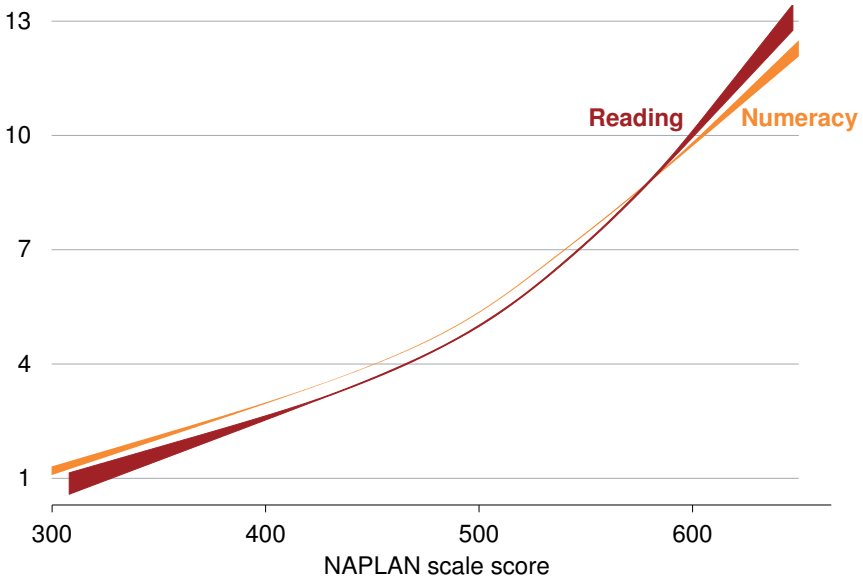
It should be noted that these confidence intervals are calculated under the assumptions in the modelling process. They tell us that the error due to measurement and sample size is likely to be small at most equivalent year levels. They do not tell us whether or not the methodology is appropriate. If we were to account for uncertain assumptions, the intervals would be wider.

[69] This assumes that individual measurement error is not systematically biased.

[70] Each replication uses a different set of random draws. The lower bound at each point is the average of the two lowest simulated points, while the upper bound at each point is the average of the two highest simulated points.

[71] In numeracy, the confidence intervals are smaller – three months at the bottom end, and five months at the top end.

Figure C.4: Confidence intervals are much wider in the extremes
Estimated 99 per cent confidence interval for equivalent year levels, Australia



Source: Grattan analysis of ACARA (2014).

Table C.1: Estimated equivalent year levels with 99 per cent confidence interval, Australia

NAPLAN score	Reading		Numeracy	
	$\widehat{Y}^*$	Interval	$\widehat{Y}^*$	Interval
325	1.30	(0.94, 1.42)	1.62	(1.55, 1.72)
350	1.74	(1.47, 1.82)	2.06	(2.01, 2.14)
375	2.17	(2.00, 2.23)	2.51	(2.48, 2.55)
400	2.61	(2.53, 2.64)	2.97	(2.95, 2.99)
425	3.08	(3.07, 3.09)	3.45	(3.44, 3.46)
450	3.62	(3.60, 3.63)	3.97	(3.96, 3.98)
475	4.25	(4.23, 4.25)	4.58	(4.57, 4.59)
500	5.01	(4.99, 5.02)	5.36	(5.34, 5.37)
525	5.98	(5.97, 6.00)	6.34	(6.32, 6.36)
550	7.16	(7.15, 7.19)	7.42	(7.40, 7.44)
575	8.51	(8.49, 8.54)	8.54	(8.53, 8.58)
600	10.00	(9.95, 10.12)	9.74	(9.71, 9.81)
625	11.57	(11.45, 11.89)	10.98	(10.89, 11.15)
650	13.15	(12.94, 13.65)	12.22	(12.08, 12.50)

Notes: Parentheses show upper and lower bounds of 99 per cent confidence interval for estimated equivalent year levels. This is estimated by a bootstrap simulation with 200 replications. Some estimated equivalent year levels and confidence bounds are below $y_{min} = 1$ or above $y_{max} = 13$, which shows how wide the intervals are at such points.
Source: Grattan analysis of ACARA (2014).

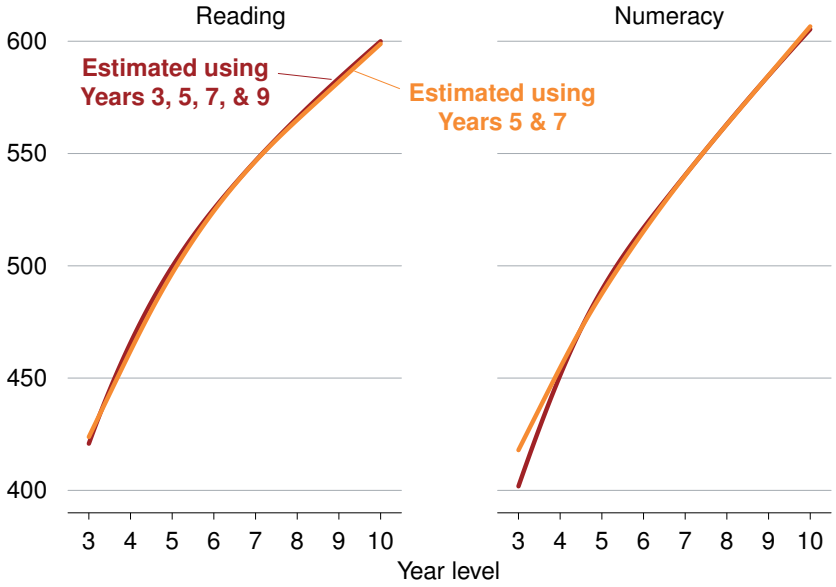
C.4.2 Accuracy of estimates beyond Year 3 and Year 9

Without students taking a NAPLAN test outside of the test-taking years, it is impossible to validate whether our estimates of the median NAPLAN scale score in Years 2, 10, and 11, for instance, reflect how the median student would actually perform in those year levels.[72] But it is possible to use a similar methodology to predict the median score in Year 3 and Year 9 without using data from Year 3 and Year 9. This can then be compared to the estimated median NAPLAN scale score for Year 3 and Year 9 on the full dataset.

Using data for students in Year 7 linked to their Year 5 results, Figure C.5 shows that the methodology predicts the median NAPLAN scale score outside these year levels with reasonable accuracy (using the curve based on the full dataset as a benchmark). There is some evidence, however, that predicting the median score for year levels well beyond the available data will lead to inaccuracies.[73]

On the whole, the results using Years 5 to 7 data provide a reasonable estimate of equivalent year levels between 18 and 24 months below Year 5, and up to two years ahead of Year 7. Although it is not possible to test the accuracy of our estimates beyond Year 3 and Year 9, these results provide some support for the robustness of the methodology.

Figure C.5: Data from Years 5 and 7 students provides a reasonable approximation for other year levels
Estimated median NAPLAN scale score, Australia



Source: Grattan analysis of ACARA (2014).

[72] As discussed in Box A.2 on page 15, equivalent year level 11 in numeracy may not actually represent the typical Year 11 numeracy student, because of curriculum changes and greater student autonomy over subject choices in senior secondary school. The issue is therefore whether equivalent year level 11 is an accurate estimate of where a typical Year 9 student would be in two years time if they continued to study numeracy or reading in a similar way.

[73] For instance, using the Years 5 to 7 data overestimates the median score in Year 3 numeracy by about 20 NAPLAN points.

C.4.3 How do estimates change with different assumptions?

Using a different benchmark student

Estimates of equivalent year levels are based on the expected path of progress of the median student. Changing the benchmark will not only change the estimated curve, $\widehat{f}_j(Y)$, but will also change the definition of the curve.

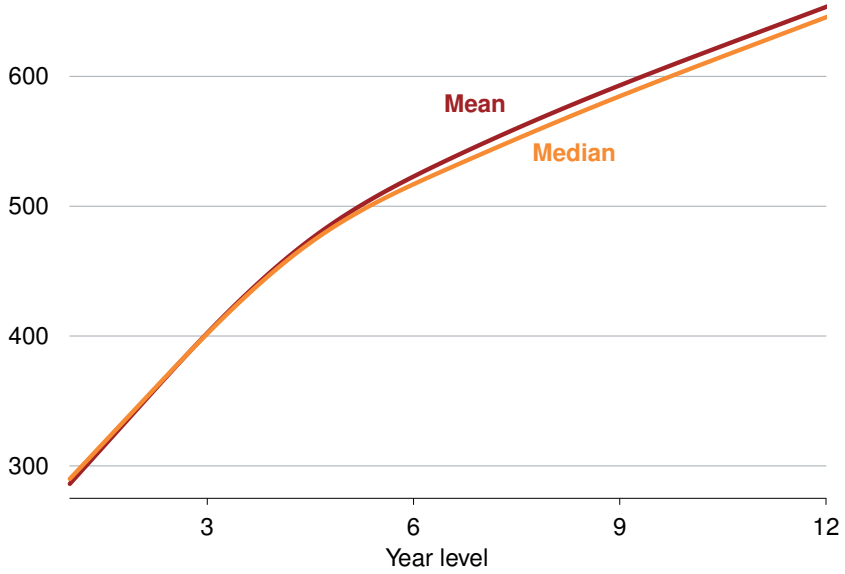
The most obvious alternative to using the median is to use the mean NAPLAN scale score in each year level. This has a noticeable, but relatively small impact on the shape of the curve, as shown in Figure C.6.[74]

Alternatively, instead of using a measure of central tendency such as the mean or median, the benchmark could be set much higher – say, at the 80th percentile. A *year of progress* would then be something harder for students to attain, but could be seen as something to aspire to. A curve based on the 80th percentile would be a better way of grouping high achieving students (for instance, those with NAPLAN scale scores above 650 in Year 9), but it would be difficult to accurately estimate what the 80th percentile student would have scored on a NAPLAN test taken before Year 3. Thus, this curve is unlikely to provide a good measure of progress over six years for average and below-average students.

In any case, it is worth noting that all percentiles between the 10th and the 90th appear to be concave, as shown in Figure C.7 on the following page. This suggests that the key findings of the report – such as the gaps in student progress between different sub-groups – would still hold even if equivalent year levels were estimated for a different percentile.

[74] This curve uses the sample means to estimate $f_j(Y)$ for $Y = 3, 5, 7$, and estimates $g_{j,Y}$ via a least squares regression.

Figure C.6: Using the mean instead of the median changes the curve slightly
Estimated median NAPLAN scale score, numeracy, Australia

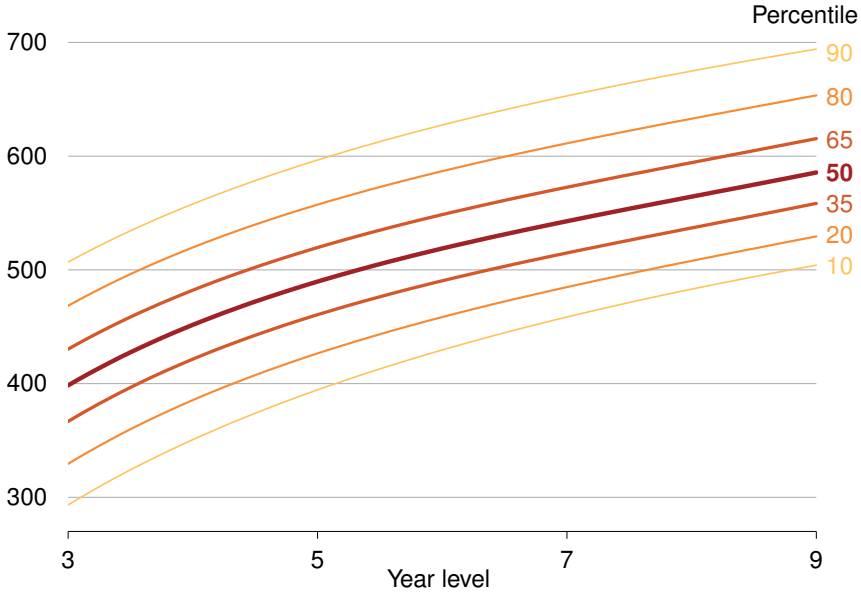


Source: Grattan analysis of ACARA (2014).

Using control variables to estimate gain scores

One assumption that was strongly considered in this methodology was to include control variables in eq. (C.6) – the equation for $g_{j,Y}$. The rationale behind this is that $\widehat{g}_{j,5}$ is estimated for below-average students, and $\widehat{g}_{j,9}$ is estimated for above-average students, even though both are used as a proxy for the median student. Including control variables such as parental education and occupation could allow us to adjust for the non-representativeness of the sample of above-average or below-average students.

Figure C.7: All percentiles make smaller gain scores at higher year levels
NAPLAN scale score by percentile, numeracy, Australia



Notes: Percentiles defined according to 2014. Each curve is smoothed across four observed points using a third-order polynomial to get a better picture of the relationship. A similar pattern occurs for reading.
Source: Grattan analysis of ACARA (2014).

This approach results in an benchmark curve that is steeper for lower scores, and flatter for higher scores. While using control variables makes intuitive sense, when $g_{j,Y}$ is estimated without control variables, our estimated equivalent year levels will provide more conservative estimates of the gaps in student progress between different sub-groups. We felt it was better to go with a more conservative approach.[75]

[75] In addition to being less conservative, using control variables may exacerbate the impact of regression to the mean, potentially introducing more error into the analysis.

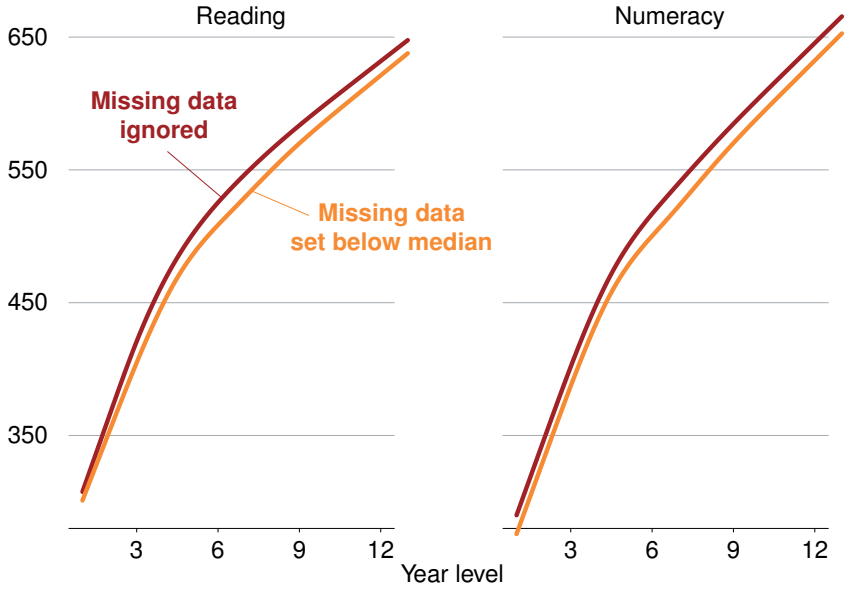
Treatment of missing data

Students that are exempt, absent, or withdrawn from a NAPLAN test in either 2012 or 2014 are ignored for the purposes of estimating the median NAPLAN scale score in each year level. But Section B.5 suggests that students who miss a test are more likely to come from households with lower parental education, and are likely to make smaller gain scores from a given prior score than other students. This means the estimated median score is likely to be above the true 50th percentile.

An alternative approach would assume that all students who missed a test would have scored below the median had they taken the test. Obviously some students that missed a test would score above the median, but it is likely that a significant majority of students who missed a test would have been below average. Thus, treating missing data as below the median may better approximate the median score than ignoring missing data.

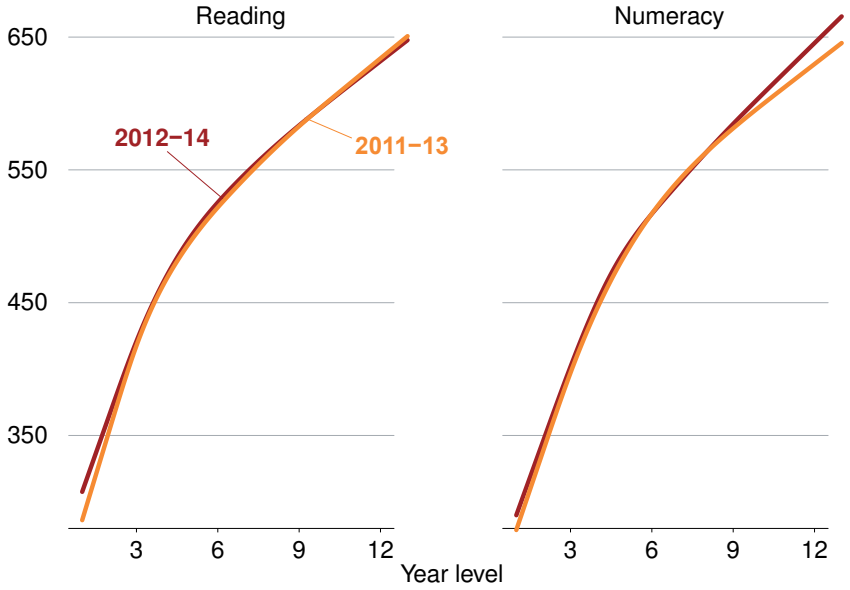
Figure C.8 shows that this alternative treatment of missing data will, unsurprisingly, lead to a lower estimate of the median NAPLAN scale score in each year level. But the curves for both reading and numeracy still have the same concave shape. It is unlikely that this alternative treatment of missing data would lead to very different conclusions about the gaps in student progress.

Figure C.8: Treating missing data as below the median does not change the shape of the curve
Estimated median NAPLAN scale score, Australia



Source: Grattan analysis of ACARA (2014).

Figure C.9: There are some discrepancies that arise with different cohorts
Estimated median NAPLAN scale score, Australia



Source: Grattan analysis of ACARA (2013) and ACARA (2014).

C.4.4 How robust are estimates to different cohorts

It is not uncommon for the distribution of NAPLAN results to change across different cohorts. This could be due to improvements or changes in the way that certain subjects are taught, or differences in the characteristics of two cohorts.[76] At the national level, results are not expected to change significantly across two cohorts one year apart.

We cross-checked our results by applying the methodology to the national cohort of 2013 students, with results linked to 2011. As Figure C.9 shows, in reading, the 2011-13 results are almost identical to those of 2012-14, except for Year 1 where the standard error is high (see Section C.4.1). In numeracy, there is little noticeable difference below Year 9, but the estimated curve using the 2011-13 data is flatter for later year levels. This means the 2012-14 numeracy curve will provide more conservative estimates of progress for high achievers, students with high levels of parental education and students from high advantaged schools.

[76] For example, Queensland introduced a Prep school year in 2008, meaning that the cohort of Year 5 students in 2013 are older than the cohort of Year 5 students in 2012.

C.5 How equivalent year levels could be implemented as part of NAPLAN reporting

Reporting NAPLAN results in terms of equivalent year levels provides a new interpretation of how students are learning relative to their peers. Given the importance of measuring student progress, and the limitations of NAPLAN gain scores, we believe this is an important contribution that should be considered as part of the official reporting of NAPLAN results by state education departments.

Of course, it is also important to consider the limitations of this approach. In terms of the methodology outlined in this chapter, equivalent year levels are not an appropriate way of reporting individual student results. This is because equivalent year levels do not cover the full range of NAPLAN scale scores, so this measure is inappropriate for high-achieving students (those performing above equivalent year level 12). In addition, high levels of measurement error at the individual level mean that it is difficult to accurately assign a student to an equivalent year level.[77]

These issues are mitigated somewhat at the school level, provided that there are a sufficient number of students to reduce measurement error, and that most students perform below Year 12 level. It should be possible to estimate an equivalent year level curve that adjusts for school background factors, but this is beyond the scope of this report. In any case, the greatest value of our approach is in measuring the progress of different cohorts and sub-groups with a common benchmark.

If this approach was to be implemented as part of NAPLAN reporting, there are a number of approaches that may improve the accuracy of the measure. First, the move to NAPLAN online will strengthen the vertical and horizontal equating process, thereby improving the accuracy of equivalent year levels. Second, it would be useful to sample students outside the NAPLAN test-taking years to validate the estimates of the median score in these years. For instance, if a NAPLAN test was given to a small number of students in Year 2 and Year 10, this would lead to more accurate estimates of performance in these year levels. Finally, the curve could be estimated as the average of multiple cohorts to reduce the discrepancies between cohorts.

[77] For a student above Year 9 standard, their standard error could easily exceed one equivalent year level.

D Tracking student progress using linked NAPLAN data

D.1 Introduction

The Victorian cohort that sat NAPLAN in Year 3 in 2009 did Year 9 NAPLAN in 2015. They provide a rich source of data to track progress over six years of schooling. Our methodology analyses this cohort in two different ways:

- by student background (parental education, school advantage, geolocation of school)
- by NAPLAN score in Year 3.[78]

When tracking results and progress, we report the progress made by the median student within each sub-group, or for the median student starting from a given percentile. While results are reported in terms of *equivalent year levels* and *years of progress*, the analysis takes place using simulated NAPLAN scale scores and gain scores; only at the very last step are these results converted to equivalent year levels.[79]

[78] We classify students according to the 20th, 50th, and 80th percentiles of Victorian performance, which we refer to as 'low, medium, and high' Year 3 score respectively.

[79] Because our results are based on percentiles, the results would not change if the simulated NAPLAN scale scores were converted to equivalent year levels before undergoing the analysis presented in this section. However, if the results were based around the *means* of different sub-groups, these would change if the simulated NAPLAN scale scores were converted to equivalent year levels before undergoing analysis.

D.2 Estimating median NAPLAN scale scores

D.2.1 By student background

For each sub-group, we estimate the NAPLAN scale score (for each of numeracy and reading) and the corresponding equivalent year level of the median student in Years 3, 5, 7, and 9. The obvious way to do this is via the sample median in each year level, but this approach could lead to progress being overstated for some sub-groups. This is because there are more missing data in higher year levels, due to greater levels of absenteeism and withdrawal, as well as students who leave Victoria. As Section B.5 shows, students from households with lower parental education typically score below average, are more likely to miss a NAPLAN test, and typically make smaller gains than other students after controlling for prior NAPLAN score. This implies that the median student that sat a particular NAPLAN test in Year 9 is likely to have scored above the observed median student when they took the test in Year 3.[80]

It is difficult to account for all the bias due to missing data, but we take an approach to estimating median scores that aims to reduce this bias. The sample median of each sub-group is used to estimate the population sub-group median in Year 3:

$$\widehat{Q_{50}[X_{j,3}|s]} = \tilde{x}_{j,3,s} \tag{D.1}$$

Where s is an indicator of sub-group.[81] This is likely to be an overestimate of the population median for Year 3, given the

[80] If all students took all tests, we would expect the median Year 3 student to match up to the median Year 9 student.

[81] See Appendix C on page 26 for explanation of notation.

patterns of missing data. But the proportion of missing data in Year 3 is relatively small, meaning that the bias is likely to be small.

For Years 5, 7, and 9, we define a function for the median sub-group NAPLAN score conditional on Year 3 score:

$$Q_{50}\left[X_{j,Y}|s, X_{j,3}\right] = h_{j,Y,s}\left(X_{j,3}\right) \quad Y = 5, 7, 9 \tag{D.2}$$

The functions $h_{j,Y,s}$ are estimated for $j =$ reading and numeracy, $Y = 5, 7, 9$, and for each subgroup using least absolute deviations. Restricted cubic regression splines are used to allow for non-linearity in $h_{j,Y,s}$. These functions are evaluated at the estimated Year 3 sample median for each sub-group, $\tilde{x}_{j,3,s}$, to estimate each sub-group population medians for Years 5, 7, and 9:

$$\widehat{Q_{50}\left[X_{j,Y}|s\right]} = \widehat{h}_{j,y,s}\left(\tilde{x}_{j,3,s}\right) \quad Y = 5, 7, 9 \tag{D.3}$$

These estimates are typically lower than the sample medians in Year 9, suggesting that this approach reduces some of the bias due to missing data.[82]

D.2.2 Estimating percentiles

We estimate the 20th, 50th, and 80th percentiles for the population in Year 3:

$$\begin{aligned} \widehat{Q_{20}\left[X_{j,3}\right]} &= \tilde{x}_{j,3}^{(20)} \\ \widehat{Q_{50}\left[X_{j,3}\right]} &= \tilde{x}_{j,3}^{(50)} \\ \widehat{Q_{80}\left[X_{j,3}\right]} &= \tilde{x}_{j,3}^{(80)} \end{aligned} \tag{D.4}$$

These are used to track progress within each sub-group (and for the population) for a given achievement level in Year 3. We estimate the median NAPLAN score in Years 5, 7, and 9 conditional on sub-group *and* Year 3 percentile:

$$\widehat{Q_{50}\left[X_{j,Y}|s, X_{j,3}\right]} = \widehat{h}_{j,y,s}\left(\tilde{x}_{j,3}^{(P)}\right) \quad Y = 5, 7, 9 \tag{D.5}$$

where P represents the Year 3 percentile, and $\widehat{h}_{j,y,s}$ has been estimated separately for each year level and sub-group.[83]

This means that for every sub-group, we have estimated median NAPLAN scale scores in Years 3, 5, 7, and 9, both for the median of the sub-group, and conditional on the Year 3 percentile for the Victorian population. Table D.1 shows these results for students who do not have a parent with a degree or diploma. Given this is a disadvantaged sub-group, the group median results are, unsurprisingly, lower than the results for the 50th percentile of the Victorian population in Year 3.

[82] This approach is still likely to overestimate the sub-group medians, since excluding missing data is likely to overstate gain scores, as evidenced by Figure B.3 on page 20.

[83] While $\widehat{h}_{j,y,s}$ is estimated separately for different sub-groups, it is not estimated separately for different percentiles.

Table D.1: For each sub-group we estimate both group medians and the medians conditional on Year 3 percentile
Estimated median NAPLAN scale score, parental education below diploma, Victorian 2009–15 cohort

	Group median	Year 3 percentile (Victorian population)		
Year level	(below diploma)	20th	50th	80th
Year 3	390.7	344.5	408.9	476.9
Year 5	477.0	452.4	487.1	526.2
Year 7	520.8	496.9	530.4	570.6
Year 9	570.6	549.4	579.3	615.3

Source: Grattan analysis of VCAA (2015).

D.3 Converting NAPLAN scale scores to equivalent year levels

Having estimated a range of NAPLAN scale scores for sub-groups, it is then possible to convert these to equivalent year levels. As outlined in Section C.2, every NAPLAN scale score within the range of the median student between Year 1 and Year 13 has a corresponding equivalent year level. Having estimated a function that maps NAPLAN scale scores onto equivalent year levels, $Y^* = \widehat{f}_j^{-1}(X_j)$, it is straightforward to find the equivalent year level corresponding to the NAPLAN scale point estimates. The reported equivalent year level includes both the schooling year and any additional months of learning.[84]

Years (and months) of progress between Years 3 and 9 for a particular sub-group is then calculated as the difference in equivalent year levels between Years 3 and 9. The median student is expected to make six years of progress over this time.

[84] We divide each year of learning into twelve months. An alternative approach that other grade-equivalent scales have taken is to divide a year of learning into ten months, noting that the school year is roughly ten months long.

D.4 Robustness of student progress results

D.4.1 Confidence intervals for student progress

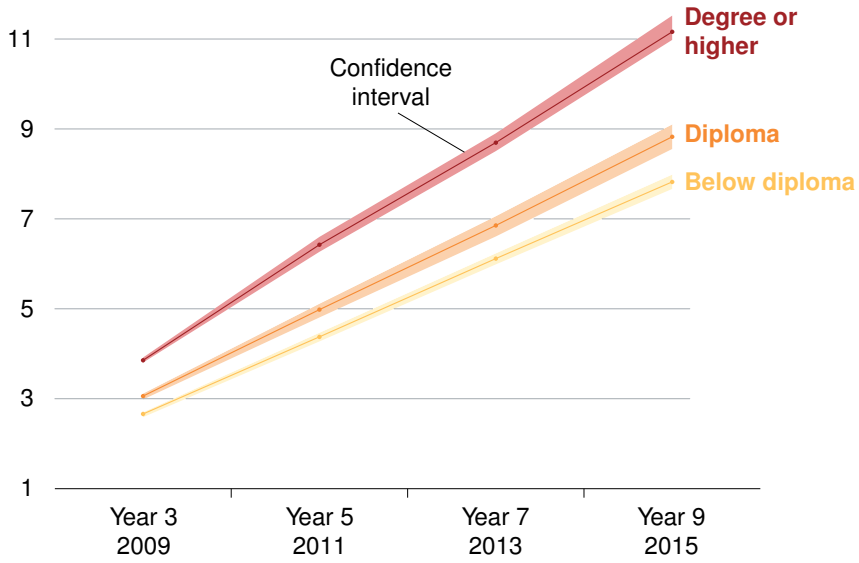
As outlined in Section B.6.3, 99 per cent confidence intervals are calculated for estimates of student progress using a bootstrap simulation. This takes into account uncertainty arising from estimation of equivalent year levels, as well as uncertainty around estimates of student progress.

Figure D.1 gives an example showing that point estimates of equivalent year levels for sub-groups are typically estimated within three months of the upper and lower confidence bounds. These are relatively narrow confidence intervals, which can be attributed to the large sample size of each sub-group analysed.

When estimating years of progress for different sub-groups from the same percentile in Year 3, statistical significance is implied by confidence intervals that do not overlap. As shown in Figure D.2, confidence intervals do not overlap for different levels of parental education from the same Year 3 score, implying that parental education is statistically significant. Full results for reading and numeracy with confidence bounds are available for download from the Grattan Institute website.[85]

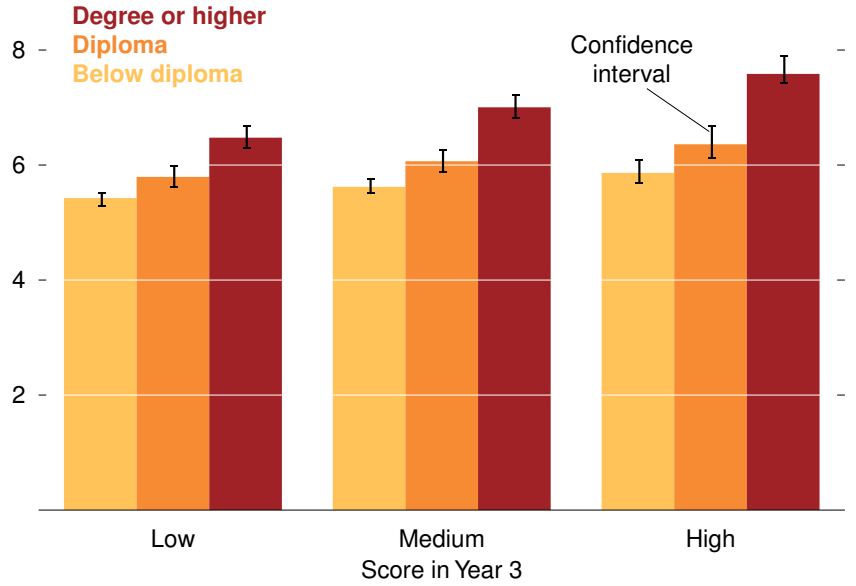
[85] http://grattan.edu.au/report/widening-gaps/

Figure D.1: The 99 per cent confidence intervals for large sub-groups are typically less than ± three months
Estimated equivalent year level by highest level of parental education, reading, Victorian 2009–15 cohort



Notes: Chart shows 99 per cent confidence interval
Source: Grattan analysis of VCAA (2015) and ACARA (2014).

Figure D.2: Confidence intervals suggest that parental education is significant in explaining student progress
Estimated years of progress by highest level of parental education, numeracy, Victorian 2009–15 cohort



Notes: 'Low, medium and high' score in Year 3 refers to the 20th, 50th and 80th percentiles for the Victorian population. Chart shows 99 per cent confidence interval
Source: Grattan analysis of VCAA (2015) and ACARA (2014).

D.4.2 Impact of regression to the mean on estimated gaps

Section B.6.2 discussed the problem of *regression to the mean*, where a student with an extreme score on one test is likely to be closer to the average score on the next test. While we were not able to correct for this in our analysis, we aimed to keep any bias arising to a minimum by avoiding extreme percentiles in our analysis.

Nonetheless, when comparing years of progress for a given Year 3 score, regression to the mean may lead to overestimation of the gaps between different sub-groups. At the 20th percentile in Year 3 (for the Victorian population), the score is more extreme for students with a university-educated parent than for those without. The Year 9 score for students with the university-educated parent will regress towards the mean score of this group, which is higher than the mean score of students without a university-educated parent. Similarly, at the 80th percentile in Year 3, the score is more extreme for students where no parent has a degree or diploma – the Year 9 score for these students will regress towards a lower mean than for other students.

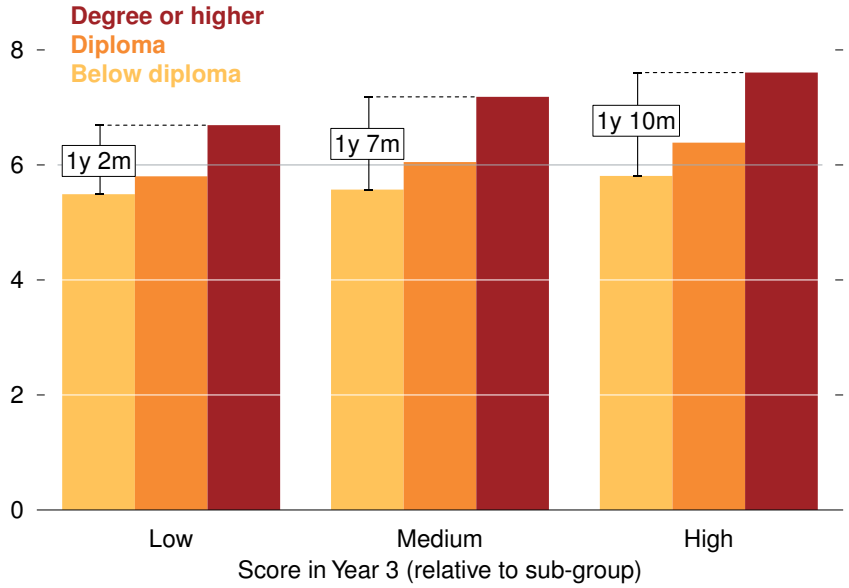
We do not believe that regression to the mean explains a significant proportion of the estimated gaps between students from households where a parent has a university degree, and those from households where no parent has a degree or diploma. For students who score at the 50th percentile in Year 3, this is relatively close to the group mean for all categories of parental education, meaning that regression to the mean will have a very small effect. Yet we still estimate a significant gap between the parental education categories 'degree or higher' and 'below diploma', consistent with the gaps found at the 20th and 80th percentiles.[86]

We also estimate years of progress conditional on the 20th, 50th and 80th percentiles *within* each parental education sub-group. Although this means we are comparing students from different starting scores, we would not expect regression to the mean to impact the estimated gaps in progress between these groups, since the within-group percentiles are as extreme as each other. As Figure D.3 shows, the gaps in progress estimated between 'degree or higher' and 'below diploma' are just as high as those estimated from the 20th, 50th and 80th percentiles for the Victorian population; they are, in fact, between one and two months of learning larger.

[86] The estimated gap is larger at the 80th percentile, and smaller at the 20th percentile.

Figure D.3: Comparing years of progress from within-group percentiles does not reduce gaps between parental education groups

Estimated years of progress by highest level of parental education, numeracy, Victorian 2009–15 cohort



Notes: 'Low, medium and high' score in Year 3 refers to the 20th, 50th and 80th percentiles within each sub-group (as opposed to Figure D.2 which is based on percentiles for the Victorian population).
Source: Grattan analysis of VCAA (2015) and ACARA (2014).

D.4.3 Robustness of results to the cohort analysed

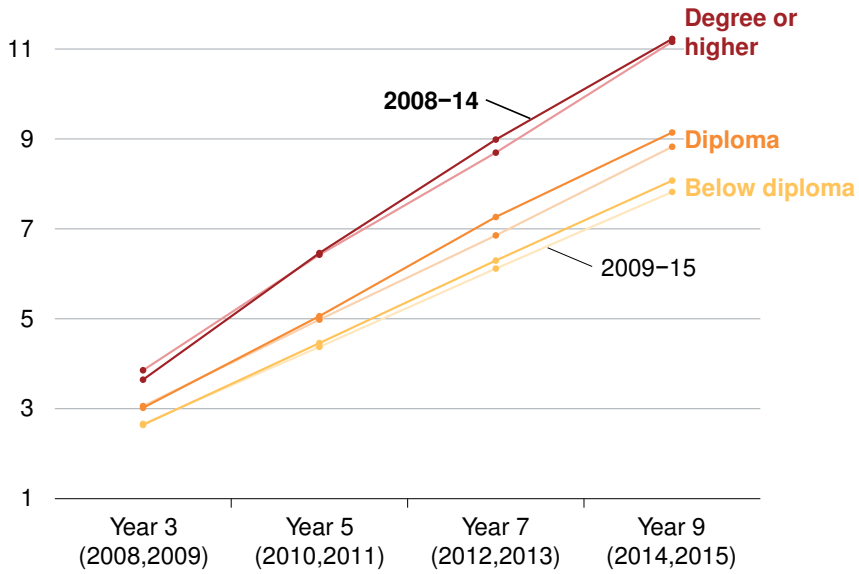
Widening gaps reports student progress based on an analysis of the cohort of Victorian students that sat the Year 3 NAPLAN tests in 2009, and the Year 9 tests in 2015. It is important to validate the results using another cohort, to see if the key conclusions are specific to this particular cohort.

We also have linked data available for the cohort of Victorian students that sat the Year 3 NAPLAN tests in 2008, and the Year 9 tests in 2014. Figures D.4 and D.5 show that the results for different sub-groups of parental education are similar across the 2008–14 and the 2009–15 cohorts, with similar gaps in student progress opening between high and low levels of parental education. However, there are differences. Year 3 students at the 20th percentile in numeracy in 2009 are estimated to make about five additional months of progress over six years as their counterparts at the 20th percentile in 2008.

We do not interpret this to mean that low achievers in the 2009 cohort made better learning progress than low achievers in the 2008 cohort. In particular, the differences may be the result of equating error or some other factor. Rather, we interpret the broad consistency of findings between the two cohorts to mean that the key findings in *Widening gaps* are sufficiently robust to inform policy decisions.[87]

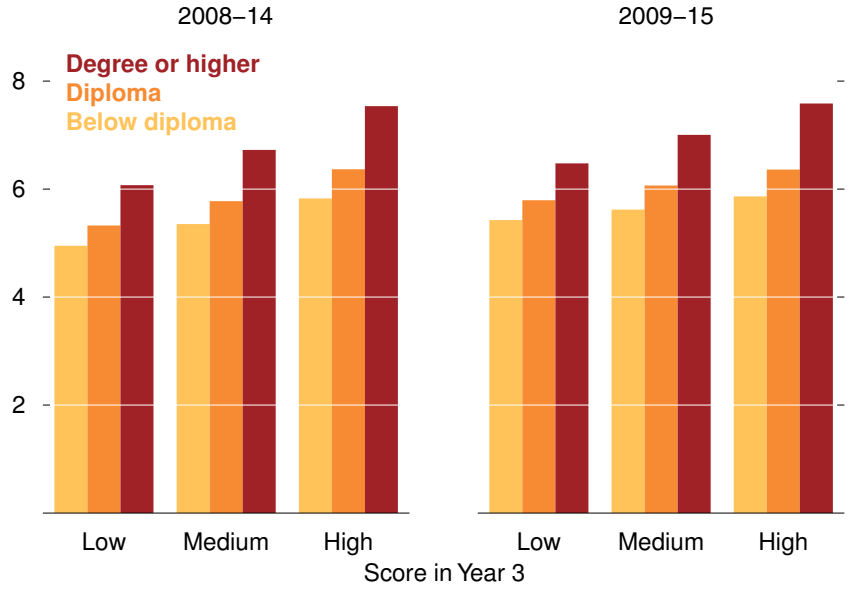
[87] Results are not shown for every sub-group analysed in the report, but the patterns of results based on the 2008–14 cohort are consistent with those for the 2009–15 cohort.

Figure D.4: Both Victorian cohorts estimate similar levels for parental education sub-groups
Estimated equivalent year level by highest year of parental education, reading, Victoria



Source: Grattan analysis of VCAA (2014), VCAA (2015) and ACARA (2014).

Figure D.5: Both Victorian cohorts estimate similar gaps in progress by parental education and Year 3 score
Estimated years of progress by highest level of parental education, numeracy, Victoria



Notes: 'Low, medium and high' score in Year 3 refers to the 20th, 50th and 80th percentiles for the Victorian population.
Source: Grattan analysis of VCAA (2014), VCAA (2015) and ACARA (2014).

Bibliography

ACARA (2013). *Deidentified student-level NAPLAN data, 2013 results linked to 2011*. Australian Curriculum Assessment and Reporting Authority, Sydney.

—— (2014). *Deidentified student-level NAPLAN data, 2014 results linked to 2012*. Australian Curriculum Assessment and Reporting Authority, Sydney.

—— (2015a). *ICSEA 2013: Technical Report*. Measurement and Research, March 2014. Australian Curriculum Assessment and Reporting Authority. http://www.acara.edu.au/verve/_resources/ICSEA_2013_Generation_Report.pdf.

—— (2015b). *My School fact sheet: Interpreting NAPLAN results*. Australian Curriculum Assessment and Reporting Authority. http://www.acara.edu.au/verve/_resources/Interpreting_NAPLAN_results_file.pdf.

—— (2015c). *NAPLAN online fact sheet*. Australian Curriculum Assessment and Reporting Authority. August 2015. http://www.nap.edu.au/verve/_resources/2015_FACT_SHEET_NAPLAN_online_tailored_tests.pdf.

—— (2015d). *NAPLAN score equivalence tables*. Australian Curriculum Assessment and Reporting Authority. http://www.nap.edu.au/results-and-reports/how-to-interpret/score-equivalence-tables.html.

—— (2015e). *National Assessment Program – Literacy and Numeracy 2014: Technical Report*. Australian Curriculum Assessment and Reporting Authority, Sydney. http://www.nap.edu.au/results-and-reports/national-reports.html.

Angoff, W. H. (1984). *Scales, Norms, and Equivalent Scores*. Princeton, New Jersey: Educational Testing Service. https://www.ets.org/Media/Research/pdf/Angoff.Scales.Norms.Equiv.Scores.pdf.

Goss, P. et al. (2016). *Widening gaps: what NAPLAN tells us about student progress*. Grattan Institute. http://www.grattan.edu.au/widening-gaps/.

Houng, B. and M. Justman (2014). *NAPLAN scores as predictors of access to higher education in Victoria*. Melbourne Institute Working Paper Series. Working Paper No. 22/14.

Marks, G. N. (2015). 'Are school-SES effects statistical artefacts? Evidence from longitudinal population data'. In: *Oxford Review of Education* 41.1, pp. 122–144.

OECD (2013). *PISA 2012 Results: Excellence through Equity: Giving every student the chance to succeed (Volume II)*. PISA, OECD Publishing. http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-II.pdf.

—— (2015). *OECD Employment Outlook 2015*. Paris: OECD Publishing.

Renaissance Learning (2015). *STAR Reading Technical Manual*. http://doc.renlearn.com/KMNet/R004384910GJF6AC.pdf.

VCAA (2014). *Deidentified linked student-level NAPLAN data, 2008 year 3 cohort*. NAPLAN results for years 3, 5, 7, and 9, 2008 to 2014. Victorian Curriculum and Assessment Authority.

—— (2015). *Deidentified linked student-level NAPLAN data, 2009 year 3 cohort*. NAPLAN results for years 3, 5, 7, and 9, 2009 to 2015. Victorian Curriculum and Assessment Authority.

Warm, T. A. (1989). 'Weighted likelihood estimation of ability in item response theory'. In: *Psychometrika* 54.3, pp. 427–450.

Wu, M. (2005). 'The role of plausible values in large-scale surveys'. In: *Studies in Educational Evaluation* 31, pp. 114–128.

—— (2010). 'Measurement, sampling, and equation errors in large-scale assessments'. In: *Educational Measurement: Issues and Practice* 29.4, pp. 15–27.