GRATTAN Institute

Closing the gaps

Technical Report

Peter Goss and Cameron Chisholm

Grattan Institute Support

Founding members

Australian Government

State Government Victoria

The University of Melbourne

bhpbilliton

Program support

Higher Education

The Myer Foundation

Affiliate Partners

Google

Origin Foundation

Senior Affiliates

EY

PwC

The Scanlon Foundation

Wesfarmers

Affiliates

Ashurst

Corrs

Deloitte

Urbis

Westpac

Grattan Institute Working Paper No 2015-13, December 2015

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, *Closing the gaps*. 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, as well as numerous industry participants and officials for their input.

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, *Closing the gaps, Technical Report*, Grattan Institute

Table of contents

List of Figures

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

A.1 Introduction

The report for Grattan Institute *Closing the gaps* seeks to measure and compare 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 to nearly all students in Years 3, 5, 7, and 9.[1] 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.[2]

These technical report includes four technical appendices to *Closing the gaps*. Appendix A describes the 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 Victorian students across Year 3 to Year 9.

A.2 The design of NAPLAN

A.2.1 NAPLAN scale scores

Students that undertake the NAPLAN test receive a score for each assessment domain: reading, writing, language conventions (which include, spelling, grammar and punctuation), and numeracy. This score, called the NAPLAN scale score, is 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.[3] Nor are the scores comparable across assessment domains.

A.2.2 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

[1] On average for a given test, about 2 per cent of students are withdrawn, 2 per cent exempt, and about 4 per cent are absent [ACARA (2014)].

[2] 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.

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

students taking the test in different years. This means that a student who took the Year 5 NAPLAN 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 test in 2013 and received the same score. That is, they are demonstrating comparable 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.[4]

While the horizontal and vertical equating process is necessary to measure student progress over time, it also introduces an additional source of error into the NAPLAN scale.[5] 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 this analysis should be revisited after NAPLAN is moved online from 2017, as online testing is likely to strengthen the equating process.[6]

[4] See ACARA (2015e), pp. 40–72 for details.
[5] See, for instance, Wu (2010).
[6] ACARA (2015c) and Wu (2010).

A.3 Looking at progress through a new lens

A.3.1 NAPLAN scale scores do not give a complete picture of student progress

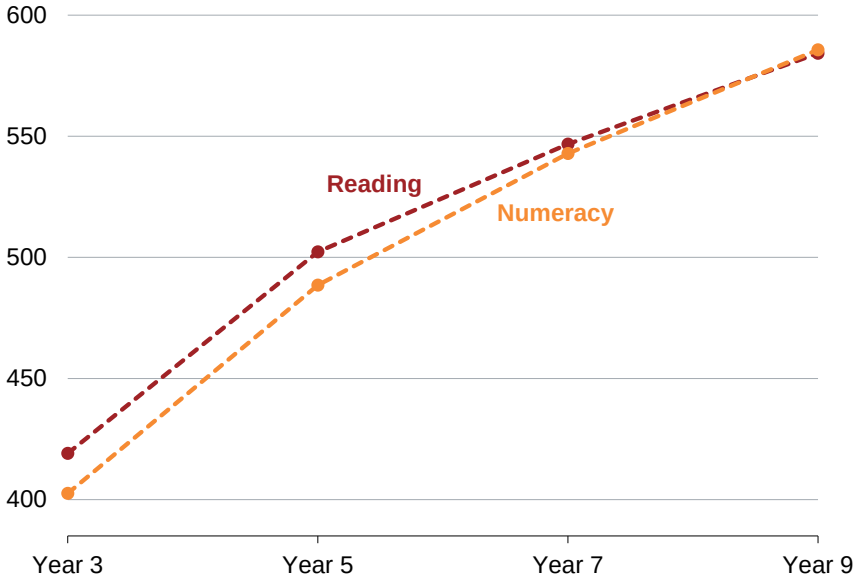
Student performance on standardised tests can be measured in a number of different ways.[7] The simplest measure, raw test scores, can be used to rank students. But raw scores can be hard to interpret. For example, on a 40-question test, the difference between 24 and 25 correct answers should not be considered equal to the difference between 38 and 39 correct answers. Raw test scores are even less useful for looking at student progress, because the measure does not take into account the degree of difficulty in the questions asked.

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.[8] This property suggests that student progress can be measured by 'gain scores': the difference between NAPLAN scale scores in two test-taking years. 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.*[9]

[7] Angoff (1984).
[8] This means that, in terms of skill level, the difference between a score of 400 and 450 is equivalent to the difference between 600 and 650, for example.
[9] ACARA (2015b).

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

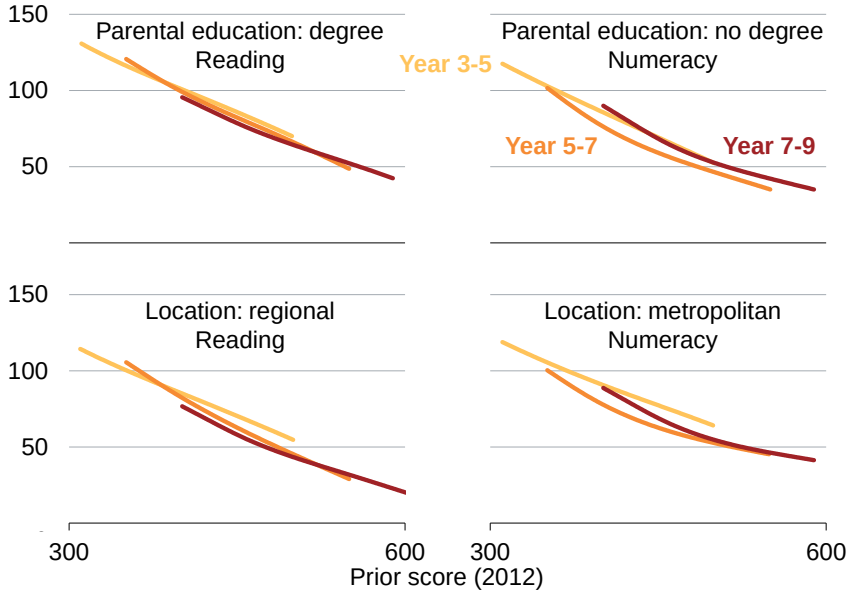


Notes: Based on 2014 and 2012 median scores.
Source: Grattan analysis of ACARA (2014).

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).[10] One interpretation of this would be to say that the education system is less effective for students in later year levels, especially between Year 7 and Year 9.

[10] Grattan analysis of ACARA (2014).

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



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).

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 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.

A third interpretation is that students genuinely learn faster from a lower base in what is being measured, and slow down over time.

Regardless of the explanation, this pattern of higher gain scores from lower starting scores should be taken into account when comparing the progress of different sub-groups of students. If not, we may draw spurious conclusions about the progress of different groups. For example, students from remote areas make higher gain scores, on average, than students from metropolitan areas between Year 3 and Year 5. 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 skill levels of remote and metropolitan students are converging. This is because remote students start from a lower skill level in Year 3.[11]

ACARA implicitly acknowledge the non-linear path of progress in the way that results are reported results against *NAPLAN proficiency bands*. There are ten proficiency bands spanning Year 3 to Year 9, with equally-spaced cut-points along the NAPLAN scale.[12] 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.[13] 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.

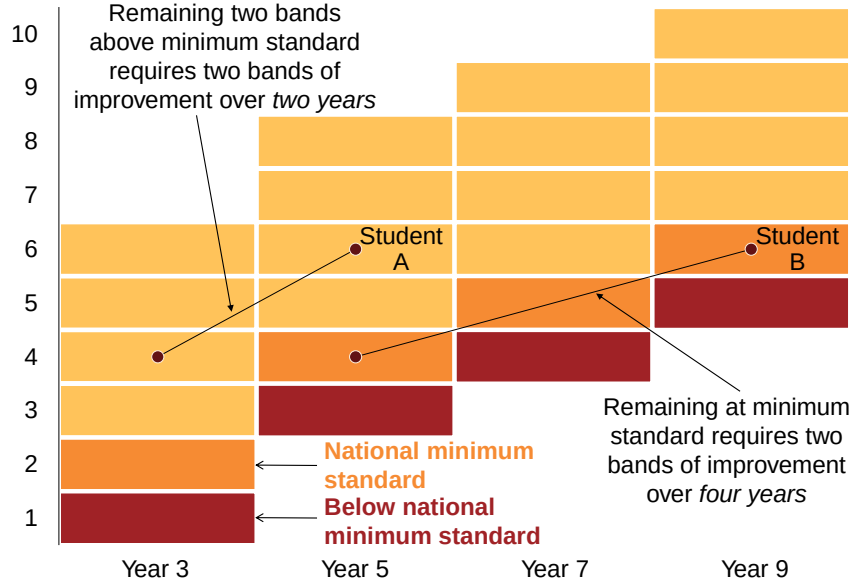
Six proficiency bands are reported for each year level. For a student to remain in the same *relative* proficiency band, they must move up two bands between Years 3 and 5, then one band between Years 5 and 7, and another band between Years 7 and 9. But students who remain in the same relative band have 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.[14]

Other researchers have accounted for non-linearity using 'like-for-like' comparisons of student progress. That is, they have only compared gain scores across different sub-groups from the same prior score.[15] This type of analysis gives a meaningful interpretation of the differences in progress between two or more sub-groups. But it is limited in its scope – many population sub-groups start from very different skill levels. To compare the progress of students starting from different skill levels requires a new lens.

[11] This example is shown in the main report, Figure 3.
[12] With the exception of Band 1 and Band 10, each band spans 52 NAPLAN scale points.
[13] 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.
[14] The analysis in this report does not use NAPLAN proficiency bands to assess student progress.
[15] It would also be valid to compare the gain scores of students who end up with the same score.

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



Source: ACARA (2015e).

A.3.2 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 is smaller than that required between Years 3 and 5.

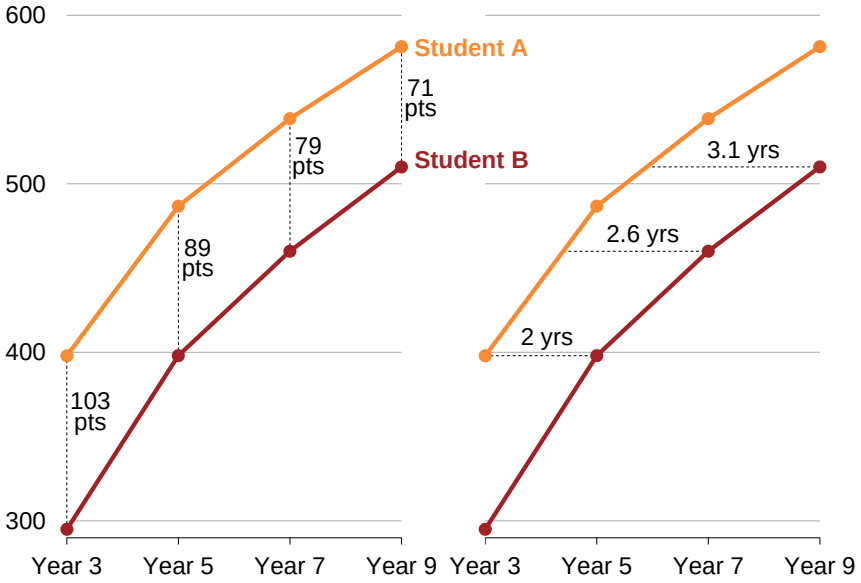
Years of progress is a measure of student progress (in terms of the skills measured by NAPLAN) relative to their peers, rather than a measure of their absolute *skill level*. This measure, as opposed to using NAPLAN gain scores, gives NAPLAN results new meaning. It can also change the interpretation of what is happening.

Consider two distinct groups of students: Group A and Group B. The scores displayed on Figure A.4 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. Looking at the gap in terms of NAPLAN points, as shown on the left chart, suggests that Group B are catching up to Group A.[16]

Yet the right chart 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 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.

[16] 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.

Figure A.4: Measuring progress in years changes the interpretation of NAPLAN results
NAPLAN scale score



Notes: The points on both charts are identical.
Source: Grattan analysis.

A.4 Measuring *Years of Progress*

If we interpret the difference between students A and B according to the right chart of Figure A.4, 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 faster 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.[17] To consistently compare progress in terms of years and months requires a reference to curriculum standards. Given that NAPLAN scores are not linked to absolute curriculum standards that define the expected capabilities for each year level, we define instead a relative benchmark.

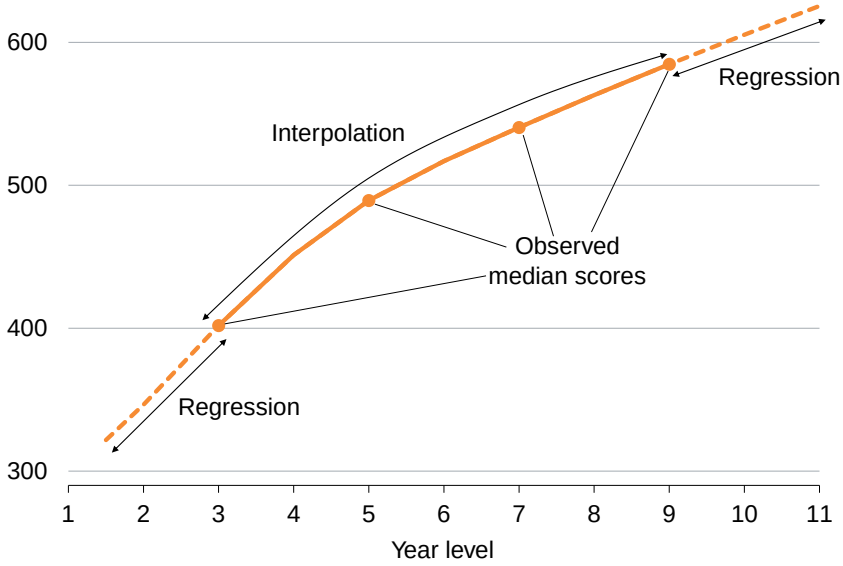
The results presented in *Closing the gaps* use the median or 'typical' student's results as a benchmark for comparing other groups of students. That is, a year of progress is defined according to the gain score expected from the median student at a given level if they were to take the NAPLAN test in one year's time.[18]

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) report a relative grade measure to compare students within each country.[19] Grade equivalent scales have also been used in reporting results of other standardised tests.[20]

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

[17] 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.
[18] 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.
[19] OECD (2013).
[20] See, for instance, Renaissance Learning (2015).

Figure A.5: Estimating equivalent year levels involves interpolation and regression
Estimated median NAPLAN scale score, numeracy



Source: Grattan analysis of ACARA (2014).

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 student learning is relative smooth at the median.[21]

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).[22] Figure A.5 shows conceptually how these approaches are used to construct a curve that maps NAPLAN scale scores to estimated equivalent year levels. The methodology is outlined in more detail in Appendix C on page 25.

[21] This might not be the case for an individual student, but is a reasonable assumption for a typical student.

[22] 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. It is possible to extrapolate the curve further than two years ahead of Year 9, but the results are less robust at such points.

Box A.1: 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 to NAPLAN gain scores.

Yet equivalent year levels can easily be misinterpreted. 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 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 these 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. 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.

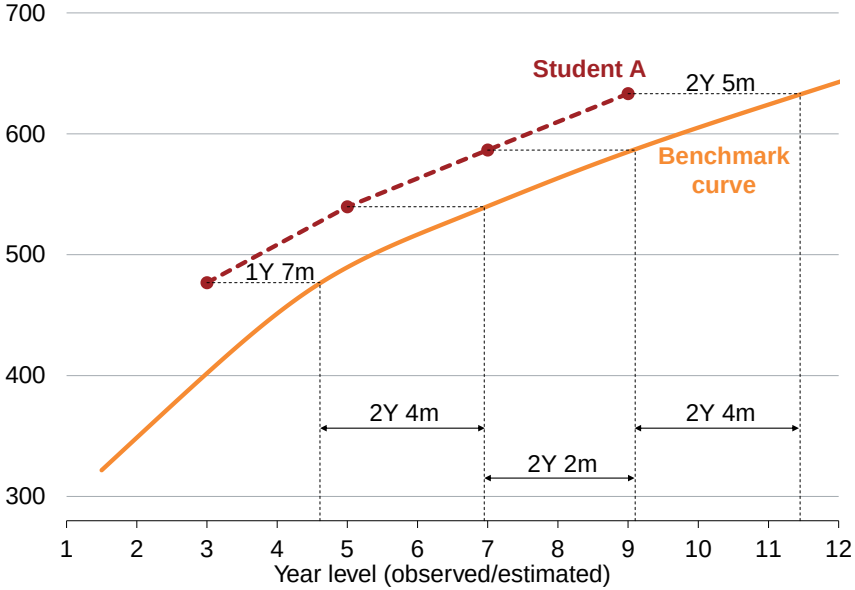
The interpretation of equivalent year levels above 9 becomes even more challenging. For instance, while we would interpret students at equivalent year level 11 in numeracy to be two years ahead of the typical Year 9 student, it is not clear whether the typical Year 9 student will reach this skill level 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 the analysis. For instance, we may only be able to look at progress between Years 5 and 7, rather than between Years 3 and 9.[a] It is much more useful to policymakers if student progress can be tracked across the majority of schooling years.

[a] In any case, our findings for student progress between Years 3 and 9 are consistent with what we find between Years 5 and 7.

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.6 for an above-average student who is about one year and seven months ahead of the benchmark curve in Year 3. By tracking this student back to the benchmark curve, we can conclude that the student made above-average progress between each NAPLAN test, finishing Year 9 two years and five months ahead of the benchmark.

Figure A.6: Student progress is measured with reference to the benchmark curve
NAPLAN scale score, numeracy



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

B Data sources and issues

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

The analysis in *Closing the gaps* is based on linked student-level NAPLAN records.[23] There are two major datasets used in the analysis:

- NAPLAN results across all four domains and year levels for all Australian students recorded in 2014, linked with their 2012 results where applicable.[24] This dataset contains test scores for more than one million students for each domain in 2014, and more than 700,000 in 2012.[25]
- NAPLAN results across all four domains recorded across 2009 to 2015 for the cohort of Victorian students who began Year 3 in 2009.[26] 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.[27]

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).[28] The national dataset contains a randomised school-level indicator – it is possible to identify whether two or more students attend the same school, but not possible to identify schools themselves.

Two additional datasets are used to check the robustness of the analysis across different cohorts – 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 across 2008 to 2014 for the cohort of Victorian students who began Year 3 in 2008.[29] Because NAPLAN results vary across cohorts, the

[23] Analysis was carried out for reading and numeracy, but not the other domains.
[24] ACARA (2014).
[25] 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.
[26] VCAA (2015).
[27] 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, about four months older than their counterparts from other states in the same year level, and are more likely to come from a high SES background [Grattan analysis of ACARA (2014)].
[28] To prevent school identification, these data were given to us in bands of 26 points.
[29] ACARA (2013) and VCAA (2014).

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.

B.2 Defining the 'typical' student

The analysis presented in *Closing the 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 SES) 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.[30] NAPLAN scale scores also tend to have a small positive skew (particularly for numeracy), which lifts the mean relative to the median.

[30] 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.

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 parental education:

- at or above Bachelor's degree
- diploma
- certificate I to IV
- year 12 or below.

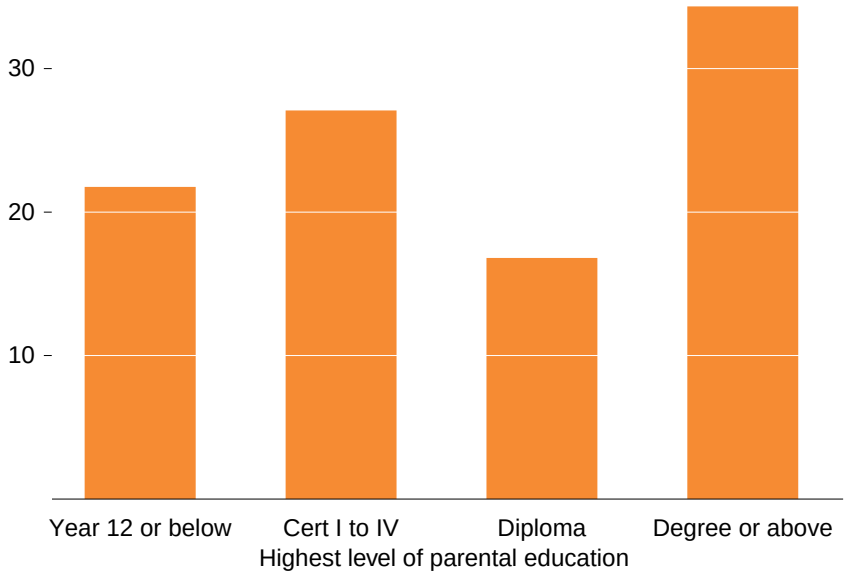
Parental education is a strong predictor of household income, and is highly correlated with other socioeconomic factors.[31]. For instance, in 85 per cent of the households where a parent has a Bachelor's degree, the highest level of parental occupation is either manager or professional, compared to only 21 per cent of households where neither parent has a degree or diploma.[32]

Figure B.1 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, we group these into a single category: 'below diploma'.

[31] See, for instance, OECD (2015)

[32] 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).

Figure B.1: Students are well represented in each category of parental education
Percentage of students, Victoria 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 SES') 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.[33]

The index is constructed as a linear combination of these 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.[34]

We use the Victorian linked data to analyse the impact of school ISCEA on student progress. Schools are allocated to one of three ICSEA groups:[35]

- ICSEA greater than 1090 (approximately the top quartile of schools in Victoria)
- 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).[36]

There is a question as to whether the strong relationship observed between school SES and NAPLAN results is

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

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

[35] 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.

[36] This cut points were chosen from the ICSEA bands available to us. It should be noted that the average ICSEA of Victorian schools is higher than the national average.

legitimate, or whether it arises as a result of the way ICSEA is constructed. 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. This means that the strong relationship observed between ICSEA and NAPLAN results is driven by SES factors, not by the way the index is constructed.

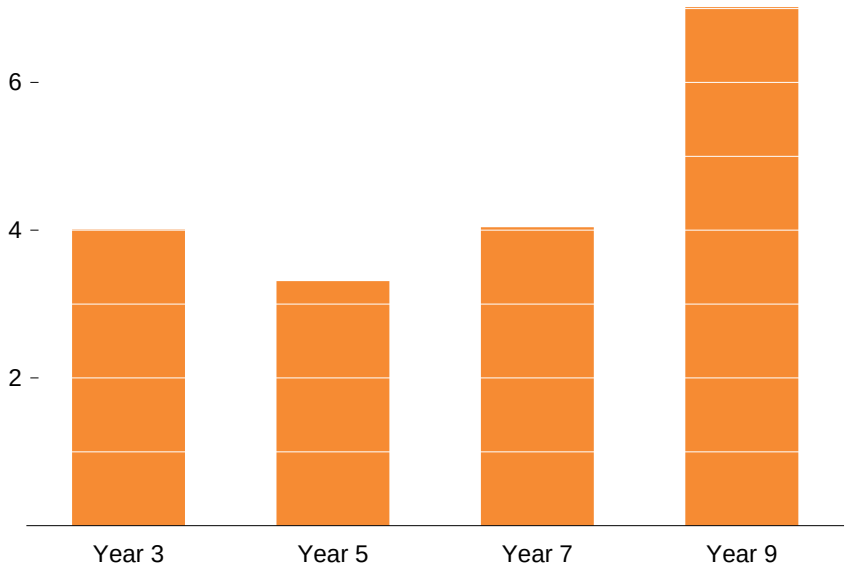
B.5 Missing data

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 different students missed the test in different years,

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



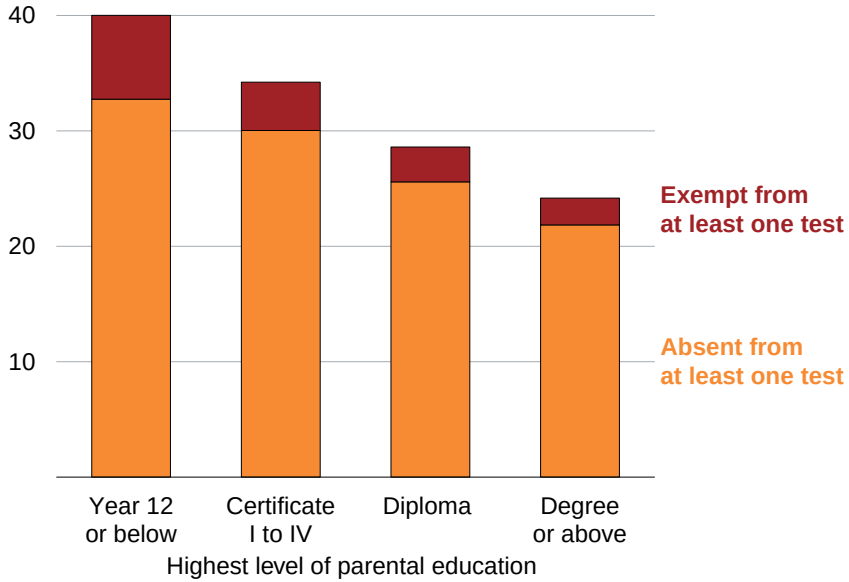
Notes: Does not include students who are exempt, withdrawn or miss a test due to leaving Victoria. Results are similar for reading.
Source: Grattan analysis of VCAA (2015).

and also because some students moved out of Victoria before Year 9.[37]

A brief analysis suggests that students are more likely to miss a test due to being absent/withdrawn or an exemption if they are from a low SES background. 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

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

Figure B.3: Students from lower SES backgrounds are more likely to miss one or more NAPLAN tests
Percentage of students that miss a NAPLAN test, Victoria 2009–15 cohort



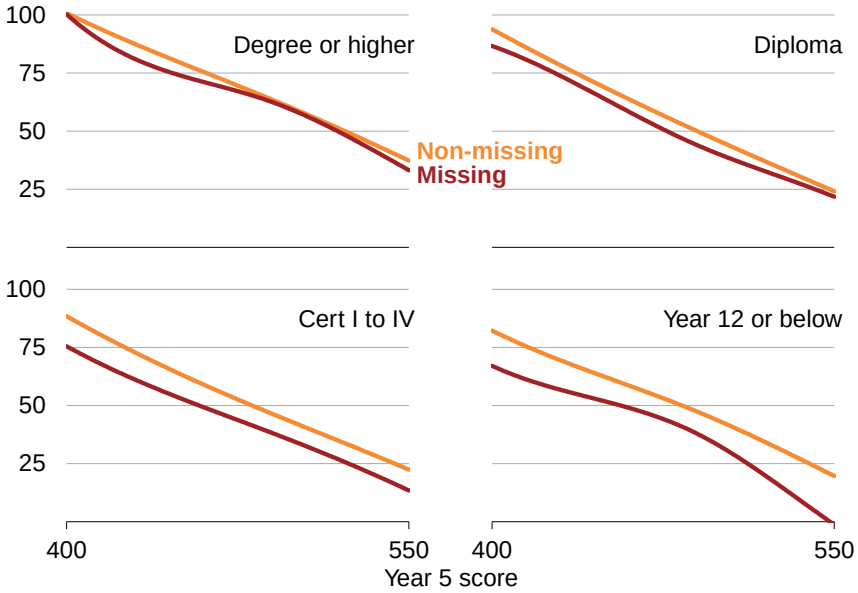
Notes: Includes all Victorian students in Year 3 in 2009, and all NAPLAN tests taken up to 2015. 'Absent from at least one test' includes those who were withdrawn, and those not in Victoria in one or more test-taking years after Year 3. Students that have been both absent and exempt from tests are categorised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

Given that students from high SES backgrounds typically score higher and make higher gains from a given starting score than those from low SES backgrounds, the consequence of ignoring missing data is an upwards bias in estimates of the median score and median gain score.[38]

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

Figure B.4: Missing data have more of an impact on gain scores for students from low SES backgrounds
Median NAPLAN gain score by highest level of parental education, reading, Year 5 to Year 7



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-level SES.
Source: Grattan analysis of VCAA (2015).

For analysis that tracks the progress of a particular sub-group, such as students that have a parent with a Bachelor's degree, missing data may be associated with lower gain scores from a given 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 shows the estimated median gain score in

reading between Year 5 and Year 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 from low SES backgrounds (using parental education as a proxy).

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 low SES groups. In turn, this means the gaps in student progress between high and low SES students are likely to be underestimated rather than overestimated.[39]

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.[40] When estimating progress for Victorian students, we take an approach that aims to minimise bias – 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.

[39] The report shows a very consistent pattern of high SES students outperforming low SES students in Year 3, and this gap growing over time. This is a key finding of the report. Missing data would be more problematic if the consequence was overestimating these gaps.

[40] 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.

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).[41] 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.[42] 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 results.

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.[43]

[41] These are also referred to as 'Warm's Estimates'; see Warm (1989).

[42] 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)].

[43] 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. This can add greater imprecision in estimating percentiles (including the median) and gain scores.

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.[44] 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.

This is best dealt with in two ways. First, adjust for the discrete nature of NAPLAN scale scores (as described in Appendix B.6.3). Second, take care when comparing cohorts across years.[45]

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/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.[46]

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.[47]

It is important to note that an extreme score for a particular sub-group might not be an extreme score for another

[44] A histogram comparing two cohorts would show a similar overall distribution, but the estimated points on the NAPLAN scale would be different.

[45] In particular, if a particular cohort is used to generate a progress baseline to assess achievement or progress of students from a different cohort, this would increase measurement error.

[46] 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 7 and Figure B.4 on page 18. 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.

[47] 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.

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 Bachelor's 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 low SES student than it is for a high achieving high SES student.[48]

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.

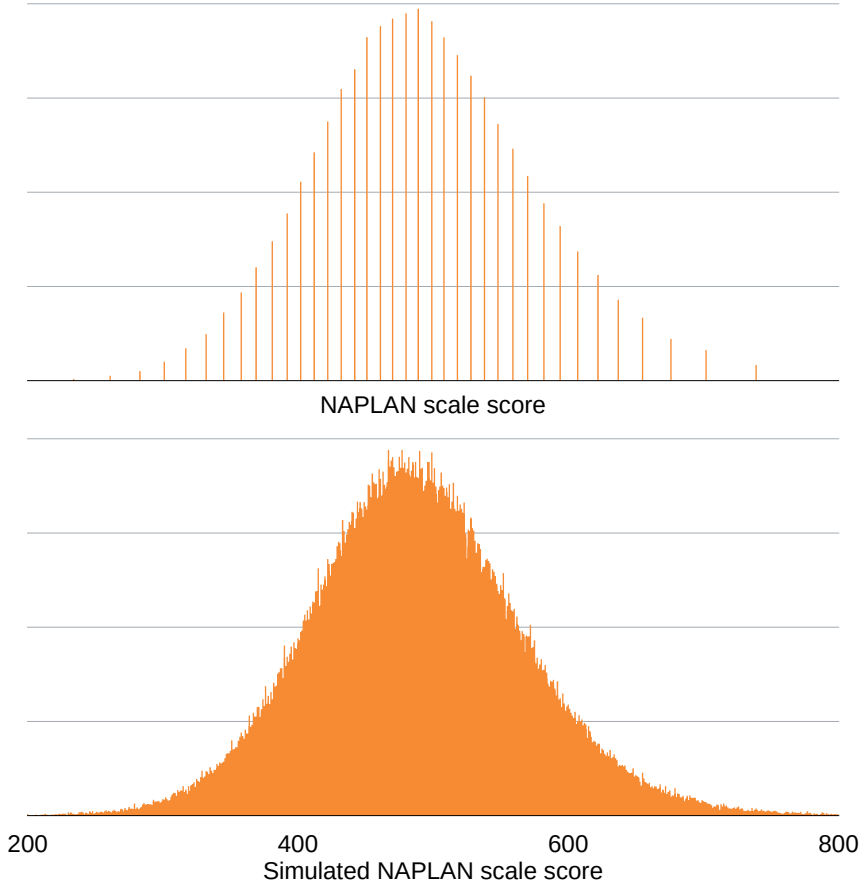
For each student, we simulate a NAPLAN scale score (skill level) as a random draw from this distribution.[49] This creates a sample that has the properties of a continuous distribution, allowing for more accurate estimates of percentiles.

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 compares a histogram of discrete NAPLAN scale scores to a histogram of simulated NAPLAN scale scores.

[48] This does not mean that all high NAPLAN scale scores for low SES students are overstating their true skill level. But when we compare a low SES and a high SES student with the same high score, the low SES student is more likely to have had a particularly good test day than the high SES student.

[49] 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).

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



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).

Use of sub-groups with large samples

Simulating NAPLAN scale 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.[50] We keep this measurement error to a minimum by aggregating students into sub-groups that have large samples, and calculating our results based on five sets of random draws.[51]

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

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

[51] 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.

estimated NAPLAN scale score or gain score of students below the 10th percentile or above the 90th percentile.[52]

In constructing the benchmark curve to estimate equivalent year levels (outlined in Appendix C on page 25), 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 six months below Year 2 level, and as high as Year 12 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 at the population level. 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 Bachelor's 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 different levels of parental education – none of these are more extreme than the 10th or 90th percentiles.

Nonetheless, the gaps in progress between high and low SES may still be overstated due to regression to the mean, particular when starting from either the 20th or the 80th percentile in Year 3. This is explore more in Section D.4.

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

Table B.1: Analysis of particular sub-groups does not extend beyond the 10th or 90th percentiles within each group
Within-group percentile in Year 3 numeracy by parental education

	Percentile		
Population	*20*	*50*	*80*
Degree or above	11.9	37.1	70.8
Diploma	19.7	50.1	81.5
Below diploma	26.3	59.5	86.7

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 or confidence bounds 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.[53] 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.[54]

[53] 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.

[54] Each replication from one simulation is linked to a replication from the other. This approach takes into account the measurement error in the

The confidence intervals are used to validate our analysis for robustness – we do not draw conclusions from any results that are not statistically significant.

Plausible values

The best approach that could be taken to reduce the impact of measurement error would be 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.[55] NAPLAN reports produced by ACARA are based on analysis using plausible values.[56]

When simulated correctly, plausible values are able to produce unbiased estimates of percentiles and gain scores for each sub-group.[57] Plausible values were, unfortunately, only 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.[58]

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.

Victorian cohort, as well as the measurement error in the estimation of equivalent year levels.

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

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

[57] Wu (2005).

[58] 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.[59] 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.

Closing the 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.[60]

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.

[59] 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.

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

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}).[61]

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:

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

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.[62]

Following this, we propose that a given NAPLAN scale score corresponds to a median schooling year – the point in time in

[61] 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.

[62] 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.

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 at each schooling year level, $Y = 1, 2, ..., 12$, 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), this would be straightforward – $\widehat{f}_j(Y)$ would just be the sample median from each year level. 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.

Step 1: Estimate the median NAPLAN scale scores 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.[63]

Step 2: Interpolate between Year 3 and Year 9

Using a quadratic function, 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.

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

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

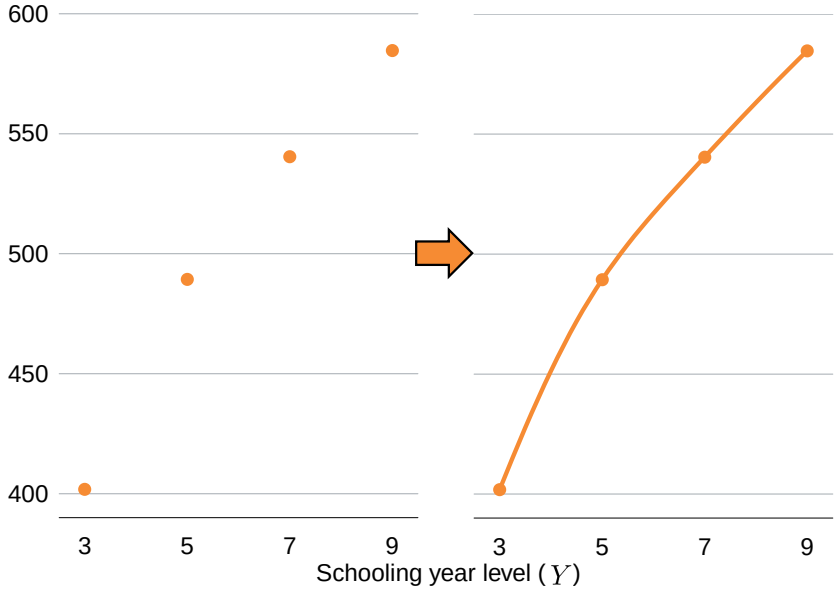
where $X_{j,Y}$ denotes NAPLAN scale score in domain j in school year Y. For students that scored $x_{j,3}$ in Year 3 reading, for example, $g_{j,5}(x_{j,3})$ is the median gain score these students will make to Year 5.[64]

[63] 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.

[64] 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.

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



Source: Grattan analysis of ACARA (2014).

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

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

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

Step 3: Estimate the median gain score curves for Years 3 to 5 and Years 7 to 9

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:

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

where S_2, S_3 and S_4 are functions that create spline variables.[65] 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.[66]

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 appears to have only a small effect on gain scores once prior scores are controlled for. For instance, when evaluated at the NAPLAN score for equivalent year level 3, $\widehat{f}_j\left(3\right)$, 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, $\widehat{f}_j\left(7\right)$, 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:

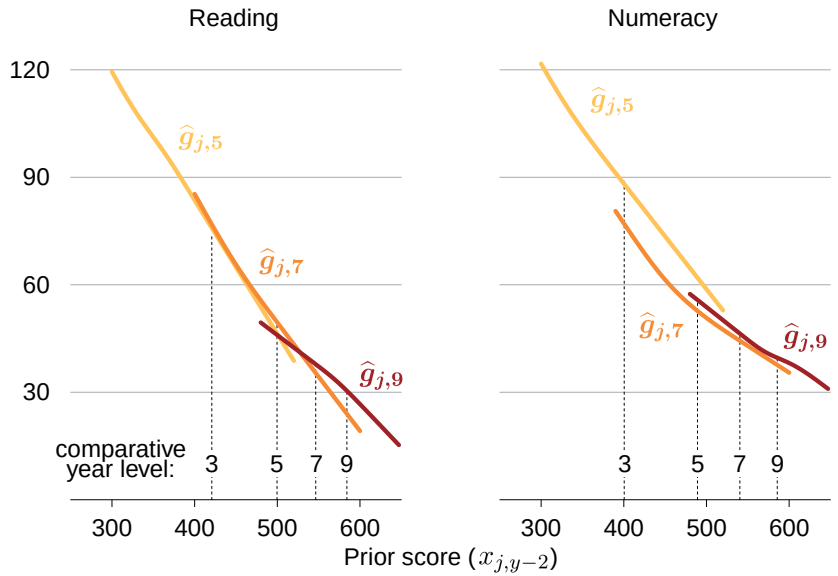
$$f_j\left(10\right) = f_j\left(8\right) + g_{j,10}\left[f_j\left(8\right)\right] \tag{C.7}$$

The point $f_j\left(8\right)$ 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

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

[66] It is only necessary to estimate $g_{j,5}$ for $x_{j,3} \leq \widehat{f}_j(3)$ and $g_{j,9}$ for $x_{j,7} \geq \widehat{f}_j(6)$.

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, $\widehat{g}_{j,y}(x_{j,y-2})$



Source: Grattan analysis of ACARA (2014).

effect on gain scores once prior scores are controlled for, we can assume:

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

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 $f_j\left(11\right)$:

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

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:

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

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

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

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 $\widehat{f}_j\left(Y\right)$:

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

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.[67] 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.[68]

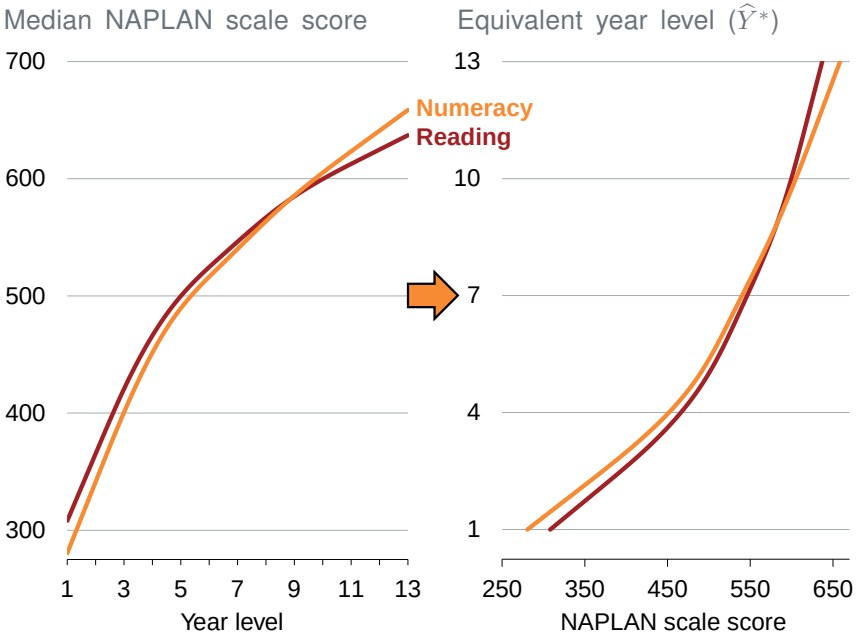
We now have a curve that estimates the median NAPLAN scale score for each schooling year level: $\widehat{f}_j(Y)$. The inverse of this curve is used to estimate the equivalent year level, Y^*, corresponding to any given NAPLAN scale score, X_j:

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

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 right chart 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.[69]

These curves can be used to compare different cohorts or sub-groups of students in terms of differences in their

Figure C.3: All NAPLAN scale scores 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).

[67] 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.

[68] See Box A.1 on page 12 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, 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

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

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. This student could be said to be learning 25 per cent faster than the typical student.

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: sample size and measurement error. But the equivalent year level curve is calculated from a very large sample, meaning that standard errors due to both sources are naturally small.

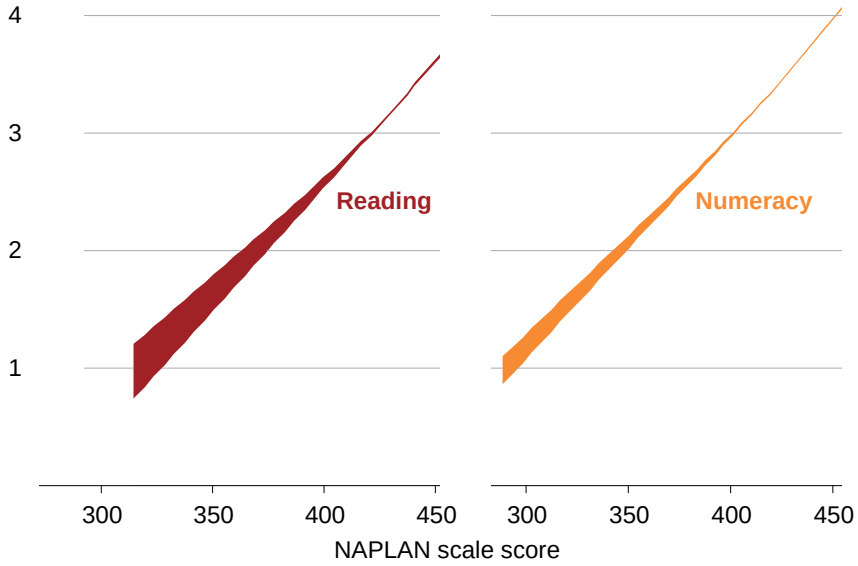
In reporting, we prefer using confidence intervals to standard errors, since equivalent year levels are asymmetrically distributed around NAPLAN scale scores. We calculate a 99 per cent confidence interval at each point along the curve, $\widehat{f}_j(Y)$, between $Y = 1$ and $Y = 13$. This is based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

Between Year 3 and Year 9, equivalent year levels are estimated within a month of learning. 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 as you go below Year 3 or above Year 9. But the interval is widest when estimating equivalent year level 1: three months for numeracy, and six months for reading, as shown in Figure C.4. At equivalent year level 13, the width of the confidence interval is two months for numeracy, and three months for reading. This reflects that there are a significant number of students who reach this level by Year 9. The confidence interval around each equivalent year level is displayed in Table C.1.

It should be noted that these confidence intervals are calculated assuming that the methodology is correct. They tell us that the error due to measurement and sample size is likely to be small. 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.

[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.

Figure C.4: Confidence intervals are widest for low NAPLAN scale scores
Estimated equivalent year level with 99 per cent confidence interval



Notes: Confidence intervals become wider for equivalent year levels greater than Year 9, but are not as wide as they are for low year levels.
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 1, 2, 10, 11, and 12 reflect how the median student would actually perform in those year levels. 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.

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

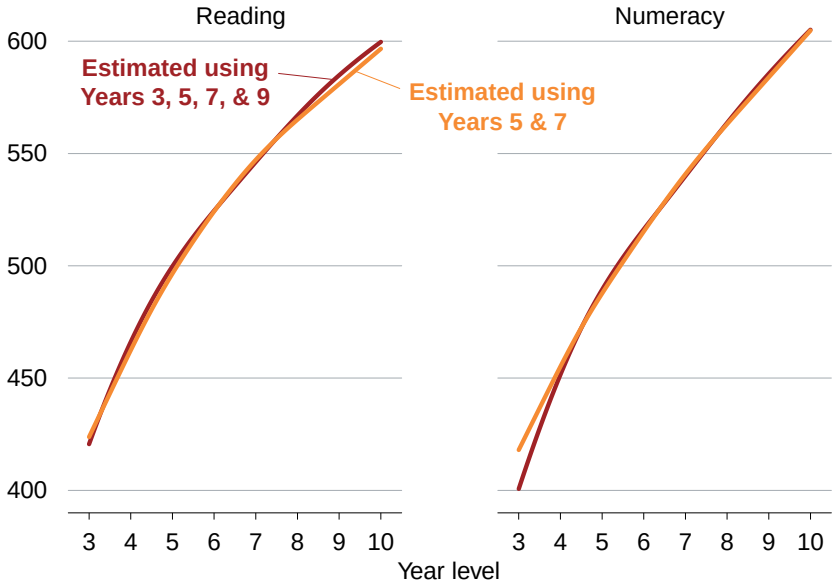
Comp. year	Numeracy		Reading	
level ($\widehat{Y}^*$)	$\widehat{f}_j(Y)$	Interval	$\widehat{f}_j(Y)$	Interval
1	288.5	(0.86,1.1)	314.8	(0.74,1.21)
2	345.4	(1.93,2.05)	368.7	(1.88,2.1)
3	401.4	(2.98,3.01)	421.1	(2.98,3.01)
4	451.4	(3.99,4.01)	466.0	(3.98,4.01)
5	489.4	(4.99,5.02)	499.9	(4.99,5.02)
6	516.2	(5.98,6.02)	524.6	(5.98,6.02)
7	540.0	(6.98,7.02)	546.2	(6.98,7.02)
8	563.9	(7.97,8.02)	566.7	(7.98,8.03)
9	586.0	(8.96,9.02)	584.6	(8.97,9.06)
10	605.6	(9.94,10.03)	599.4	(9.97,10.08)
11	623.8	(10.93,11.04)	612.4	(10.95,11.09)
12	641.4	(11.92,12.05)	624.8	(11.92,12.11)
13	659.0	(12.9,13.06)	637.1	(12.88,13.13)

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.
Source: Grattan analysis of ACARA (2014).

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.[71]

[71] For instance, using the Years 5 to 7 data overestimates the median score in Year 3 numeracy, and underestimates the median score in Year 10 reading.

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



Source: Grattan analysis of ACARA (2014).

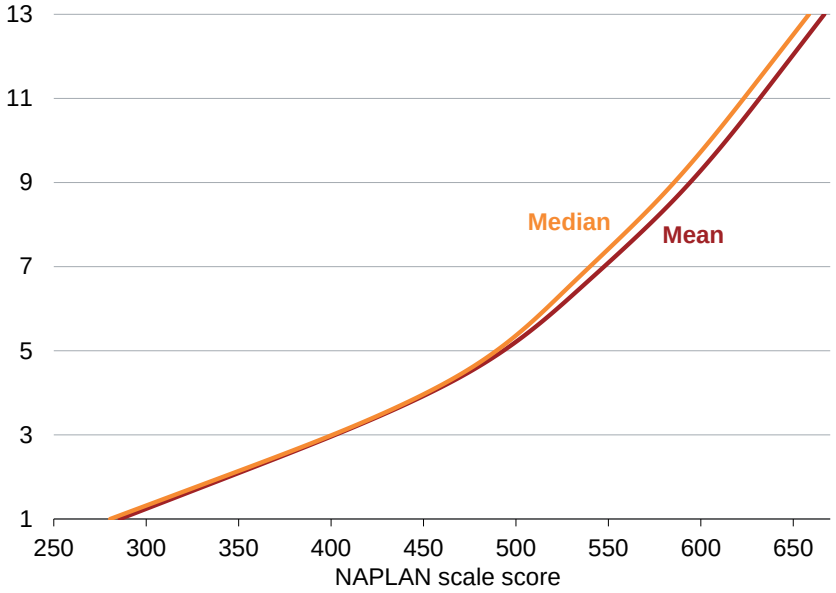
On the whole, the results using Years 5 to 7 data suggest that it is possible to estimate NAPLAN scale scores beyond Year 3 and Year 9 with a degree of accuracy.

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 a typical student, which is defined as 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.

Figure C.6: Using the mean instead of the median changes the curve slightly
Estimated equivalent year level, numeracy



Source: Grattan analysis of ACARA (2014).

The most obvious alternative to using the median is to use the mean NAPLAN scale score in each year level. This has a significant, but relatively small impact on the shape of the curve, as shown in Figure C.6.[72]

Alternatively, instead of using a measure of central tendency, 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

[72] 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.

NAPLAN scale scores about 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.

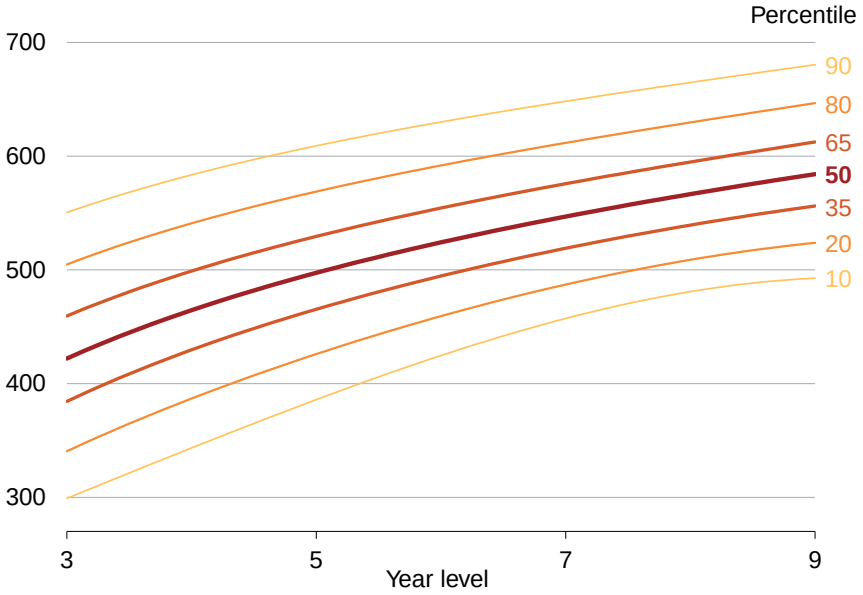
In any case, it is worth noting that all percentiles between the 10th and the 90th appear to be concave, as shown in Figure C.7. This suggests that the key findings of the report – such as the gaps in student progress between different sub-groups – will still hold even if equivalent year levels are estimated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

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.

This approach results in an estimated equivalent year level curve, $\widehat{f}_j^{-1}(X_j)$, that is much flatter for lower scores, and even steeper 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.

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



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 numeracy.
Source: Grattan analysis of ACARA (2014).

Treatment of missing data

Students that are exempt, absent, or withdrawn with 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 disadvantaged backgrounds, and are likely to make smaller gain scores from a given prior score than other students. This means that estimated median score is likely to be above the true 50th percentile.

An alternative approach would be to 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 scored below the median. 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, meaning that the conclusions regarding the gaps in student progress for different sub-groups are likely to remain the same.

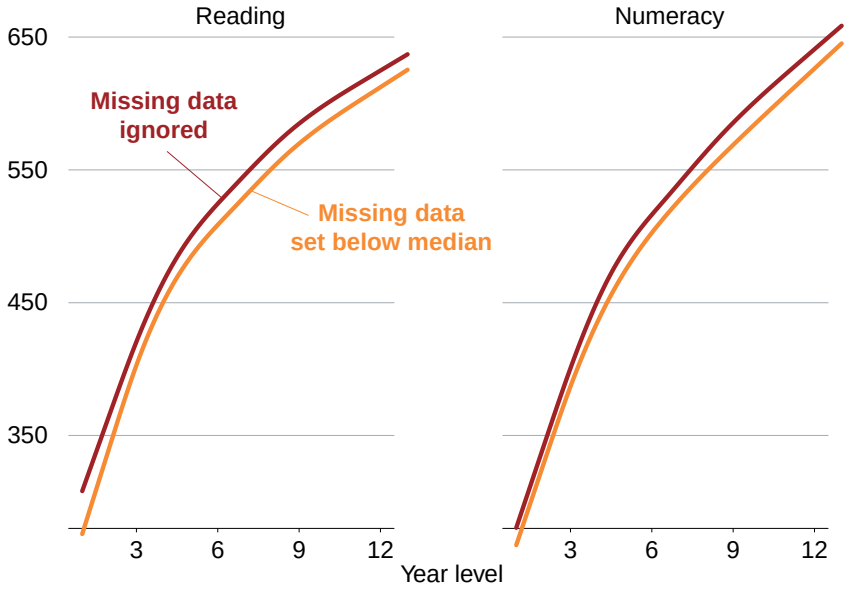
C.4.4 How robust are estimates to different cohorts

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.[73] At the national level, results are unexpected 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 Appendix C.4.1). In numeracy, there is no noticeable difference below Year 9, but the estimated curve using the 2011-13 data is flatter for later year levels.

[73] 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.

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



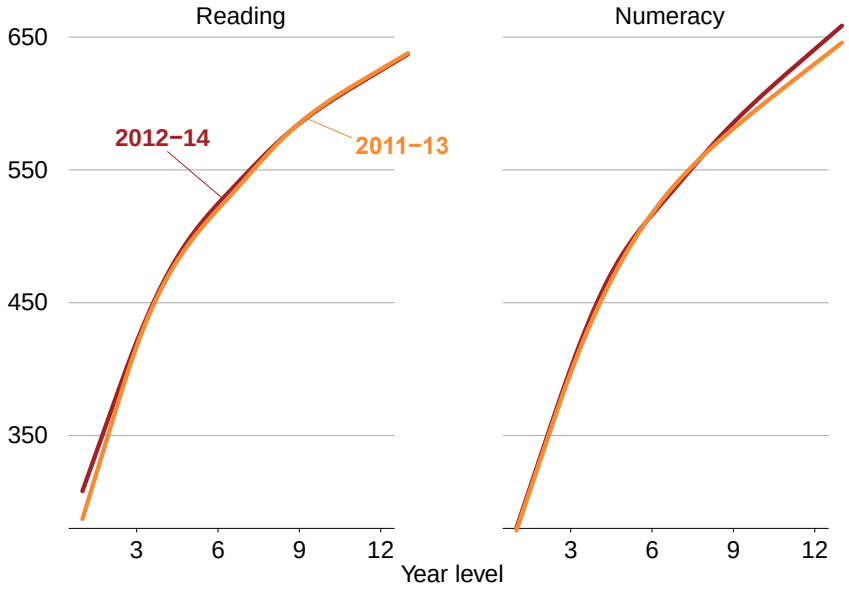
Source: Grattan analysis of ACARA (2014).

In both cases, the 2012-14 curve will provide more conservative estimates of student progress for high achievers and high SES sub-groups, and less conservative estimates of student progress for low achievers and low SES sub-groups.

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,

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



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

we believe this is an important contribution that should be considered as part of the official reporting of NAPLAN results.

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 Year 12 level). 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.[74]

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 may 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.

If this approach was to implemented as part of NAPLAN reporting, there are a number of approaches that may improve the accuracy of the measure. Firstly, 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 the test was given to a small number of students in Year 1 and Year 11, this would lead to more accurate estimates of performance in these year levels. Secondly, the curve could be estimated using multiple cohorts to reduce the discrepancies between cohorts.

[74] 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 (household SES, school SES, geolocation of school)
- by NAPLAN score in Year 3.[75]

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

D.2 Estimating median NAPLAN scale scores

D.2.1 SES sub-groups

For each SES 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 who miss one or more tests are more likely to come from lower SES backgrounds, and typically make smaller gains than other students even after controlling for SES background and prior NAPLAN score. This means the median student that sat NAPLAN in Year 9 is likely to have been above the median student in Year 3.[76]

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:

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

Where s is an indicator of sub-group.[77] This is likely to be an overestimate of the population median for Year 3, given the 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

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

[76] If students are tracked consistently, then the median Year 3 student is expected to match up to the median Year 9 student.

[77] See Appendix C on page 25 for explanation of notation.

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:

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

These estimates are typically lower than the sample medians at higher year levels, suggesting that this approach reduces some of the bias due to missing data.[78]

D.2.2 Estimating percentiles

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

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

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

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

where P represents the Year 3 percentile, and $\widehat{h}_{j,y,s}$ has been estimated separately for each year level and sub-group.[79]

[78] 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 18.

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

This means that for every SES 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 low SES sub-group, the group median results are, unsurprisingly, lower than the results for those at the 50th percentile of the Victorian population in Year 3.

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

		Year 3 percentile (Victorian population)		
Year level	Group median (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 SES 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 $\widehat{f}_j(Y)$, it is straightforward to find the value of Y that corresponds to a given $\widehat{f}_j(Y)$. The reported equivalent year

level includes both the schooling year and any additional months of learning.

Years of progress between Years 3 and 9 is then calculated as the difference in equivalent year levels between Years 3 and 9. This is also reported in terms of years and months. The median student is expected to make six years of progress over this time.

D.4 Robustness of student progress results

D.4.1 Confidence intervals for student progress

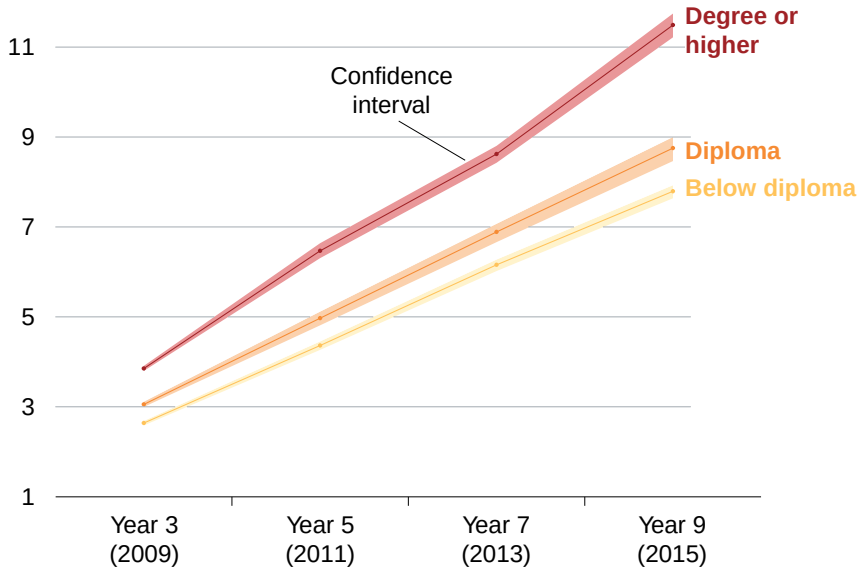
As outlined in Section B.6.3, confidence intervals are calculated for estimates of student progress using a bootstrap simulation. This takes into account uncertainty arising from estimation of equivalent year levels, as well as uncertainty around estimates of student progress.

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 with confidence bounds are available for download from the Grattan Institute website.[80]

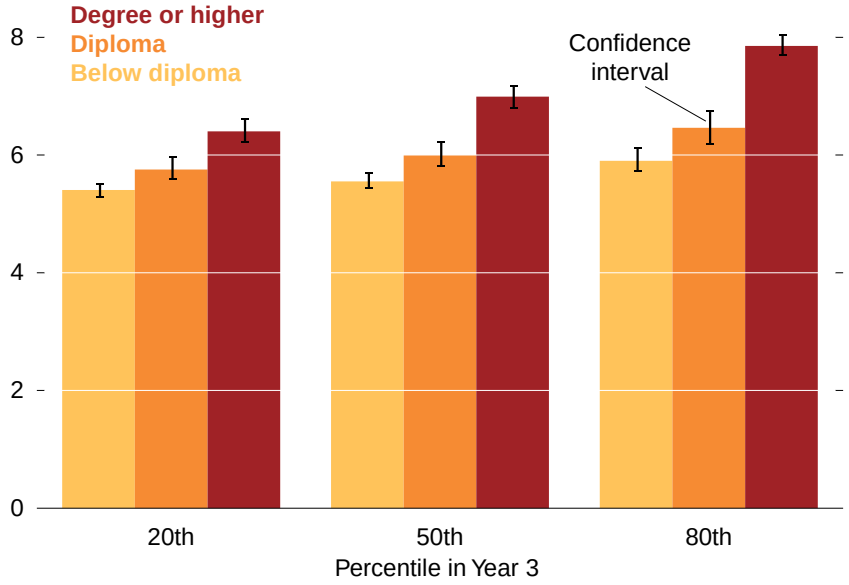
Figure D.1: Equivalent year levels for large sub-groups are typically estimated within three months of confidence bounds
Estimated equivalent year level by highest level of parental education



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

[80] http://www.grattan.edu.au/

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



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

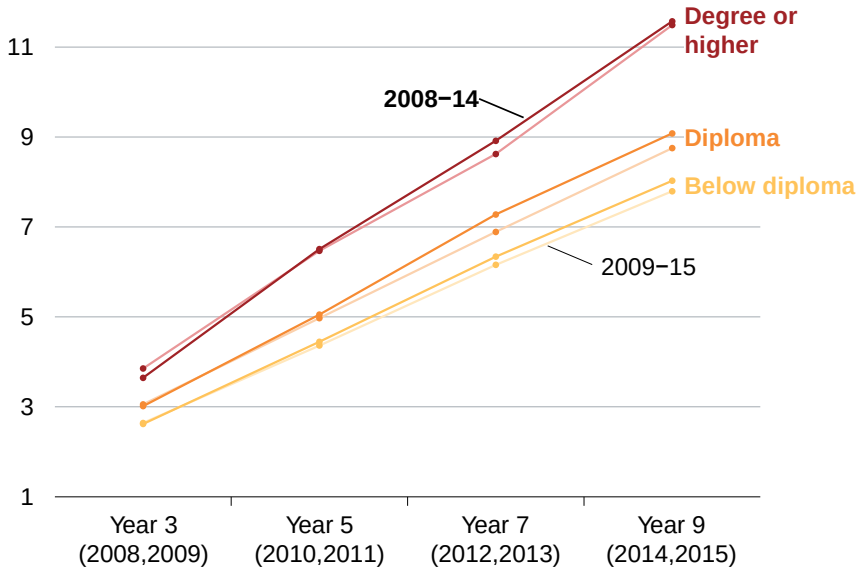
D.4.2 Robustness of results to cohort analysed

Closing the 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. Some of the key conclusions may be specific to this particular cohort, so it is important to validate the results using another 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.3 and D.4 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. 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.

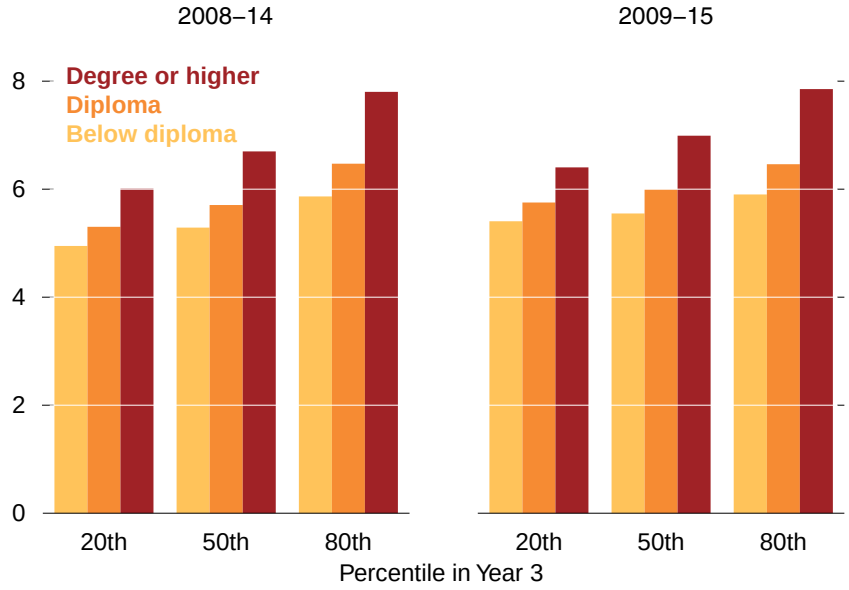
Results are not shown for every sub-group analysed in the report, but the results based on the 2008–14 cohort are consistent with those for 2009–15.

Figure D.3: Both Victorian cohorts estimate similar levels for parental education sub-groups
Estimated equivalent year level by highest year of parental education



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

Figure D.4: 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



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.

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.