How should student progress be measured on the NAPLAN test?

Cameron Chisholm* and Peter Goss†

1st April 2016

Abstract

One of the benefits of the NAPLAN test is the ability to track student performance over time. But when comparing the progress made by different groups of students, it is inappropriate to use the NAPLAN scale unless comparing students from the same starting score. This paper proposes a time-based measure of student progress on NAPLAN, *years of learning progress*, that does not have this restriction. This measure of student progress, compared to the existing scale, leads to very different conclusions about which student groups are performing well.

JEL classification: I21, C02, C23

Keywords: NAPLAN, Student progress, Equivalent year level, Years of learning progress

1 Introduction

Many education systems use standardised tests. While these can be used to identify struggling students and compare the performance of schools, standardised tests are a particularly useful tool for policymakers and researchers. When aggregated over a cohort of students, test results help to identify the characteristics of schools and students that may require additional education funding, and can also be used to test the effectiveness of policy interventions.

Standardised tests that are administered to the same students at different stages of their schooling lives can be used to measure student progress. This allows researchers to

* Grattan Institute, Melbourne, email: cameron.chisholm@grattan.edu.au
† Grattan Institute, Melbourne

look at how well early achievement predicts the rate at which students learn through school. Do low achievers learn as much as high achievers, or do they fall further behind? Do early achievers get further ahead, or are they being prevented from reaching their potential by coursework that is too easy? An accurate and meaningful measure of student progress is essential to answer such questions.

In Australia, the National Assessment Program – Literacy and Numeracy, or NAPLAN, is a standardised test administered to students in Years 3, 5, 7, and 9 across four assessment domains. A defining feature of NAPLAN is that the scale scores received by students are comparable across year levels. Because of this property, a logical way to measure progress between two points in time is to look at the difference in NAPLAN scale scores ('gain scores'). But such an approach can lead to non-meaningful and even inaccurate conclusions being drawn.

To illustrate this, picture a cycling race which starts with a flat bit of road followed by a mountain climb. The cyclist in the lead has already started the climb, and is currently riding at 20 kilometres per hour. A cyclist towards the back of the pack is still on the flat and is currently riding at 30 kilometres per hour. Is this cyclist catching up to the leader? In terms of the distance between them, the answer is unequivocally yes, but this is not a meaningful comparison. Eventually this cyclist will also reach the climb and is likely to slow down and fall further behind the leader.

For this reason, the sport of cycling reports a time gap between cyclists, usually measured in minutes and seconds, rather than the distance between cyclists. This paper proposes using a time gap to compare the progress of different groups of students on the NAPLAN test, measured in years and months. This is because making gains on the NAPLAN test is like riding up an ever-increasing slope – it's easy to make large gain scores from a low base, but difficult from a high base.

The time-based measure proposed, *Years of learning progress*, was used in a report published in March 2016 for Grattan Institute, *Widening gaps: what NAPLAN tells us about student progress*.[1] This report identified significant gaps in student progress between different groups of students, such as between those at advantaged schools and those at disadvantaged schools. These gaps could not be identified using the NAPLAN scale.

This paper is outlined as follows: Section 2 looks at the design of the NAPLAN scale and its limitations for measuring student progress. Section 3 outlines the rationale for and approach to constructing a time-based measure of student progress. Section 4 provides some student progress results obtained using this measure, showing how these can be significantly different to those measured with the NAPLAN scale. Section 5 suggests some extensions to this approach. Two appendices accompany this paper: Appendix A

[1] Goss et al. (2016).

outlines some of the data and methodological issues with NAPLAN, and Appendix B provides a detailed description of how *years of learning progress* is constructed.

2 The design of NAPLAN

2.1 NAPLAN scale scores

Students that undertake the NAPLAN test receive a score for each assessment domain: reading, writing, language conventions (which includes spelling, grammar and punctuation), and numeracy. This score, called the NAPLAN scale score, is typically between 0 and 1000. While the scores are used to indicate whether a student is above NAPLAN national minimum standards for each year level, they have no other direct interpretation. The scores are an estimate of student skill level at a point in time, a latent concept – the numbers themselves have no particular meaning.[2] Nor are the scores comparable across assessment domains.

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

The horizontal equating process involves a sample of students taking an equating test in addition to the NAPLAN tests. A scaling process takes place using this equating sample and common items across years on the equating tests. The result is that NAPLAN scale scores are comparable across different years. The vertical equating process involves common test items on the tests administered to different year levels. The results are scaled so that scale scores are comparable across different year levels.[3]

[2] It would be possible to link NAPLAN scale scores to curriculum standards, but this has not yet been developed.

[3] See ACARA (2015d), pp. 40–72 for details.

While the horizontal and vertical equating process is necessary to measure student progress over time, it also introduces an additional source of error into NAPLAN results.[4] We note that any errors arising from the equating process reduce the reliability of the *years of learning progress* measure. We intend to revisit this analysis after NAPLAN is moved online from 2017, as online testing is likely to strengthen the equating process.[5]

2.3 NAPLAN scale scores give an incomplete picture of student progress

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

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

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

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

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

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

[5] ACARA (2015b) and Wu (2010).

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

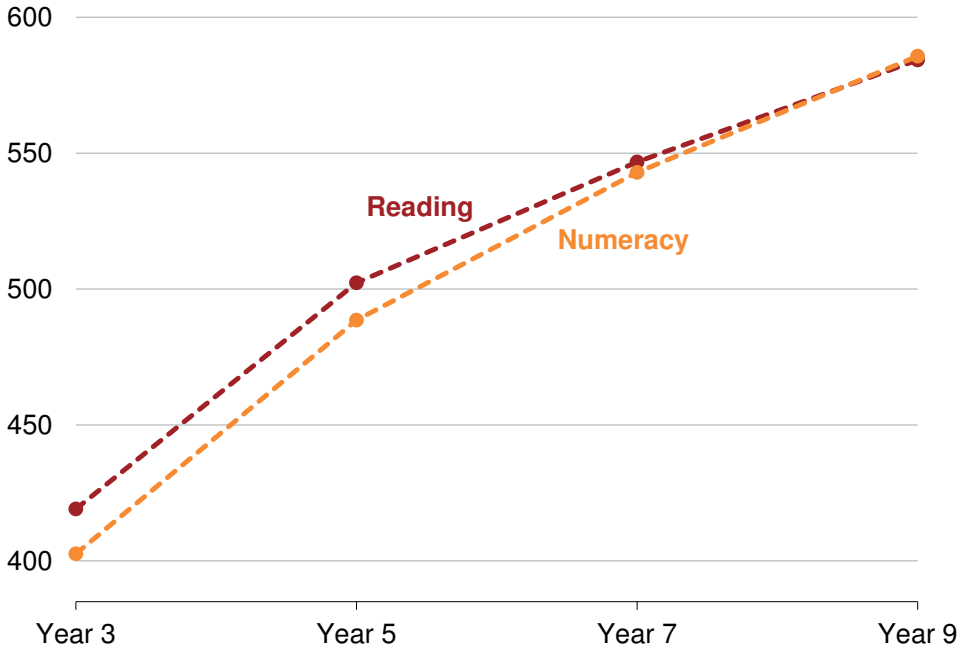
[7] NAPLAN is a test of specific literacy and numeracy skills. These skills are fundamental to student learning. Yet a standardised test does not cover all elements of student learning; for instance, NAPLAN tends to focus on specific skills rather than content knowledge. Thus, when the report refers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

[8] ACARA (2015a).

[9] Analysis of ACARA (2014).

Figure 1: The relationship between NAPLAN scale scores and year level is not linear for the median student

NAPLAN scale score of median student in each year level, Australia



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

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 2 on the next page suggests that this is not the case: lower prior scores are associated with higher gain scores *within* each year level, and the same pattern holds for different population sub-groups.[10]

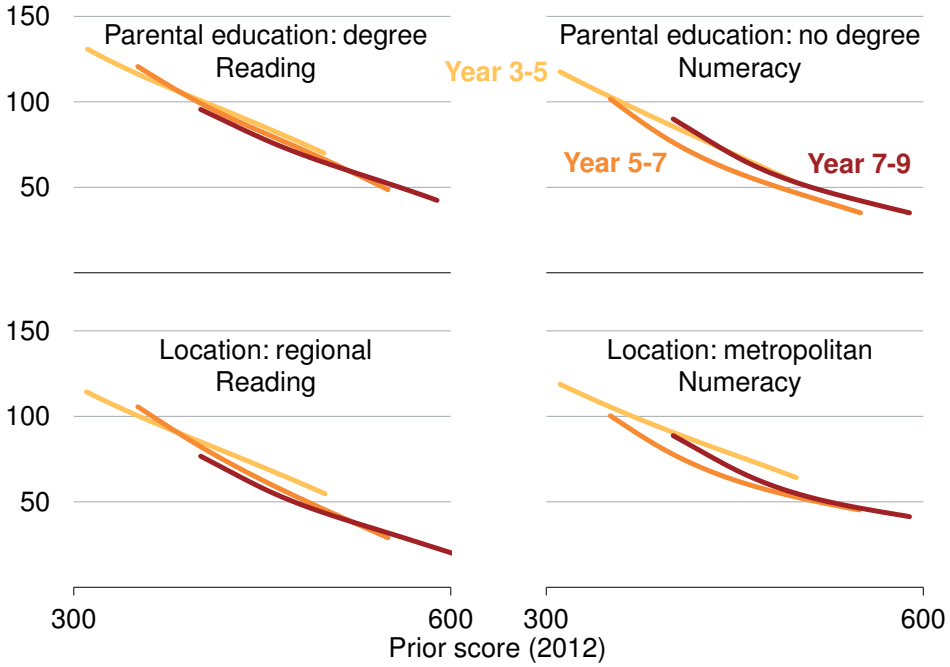
A third interpretation is that students genuinely increase their skill level faster from a lower base, and slow down over time. That is, the higher a student's current skill level, the longer it takes to increase their skill level by a given amount (as measured by the NAPLAN scale). This appears to be the favoured interpretation among psychometricians.

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

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

Figure 2: Higher gain scores are observed for lower prior scores, regardless of year level or population sub-group

Median NAPLAN gain score over two years by prior score, 2014, Australia

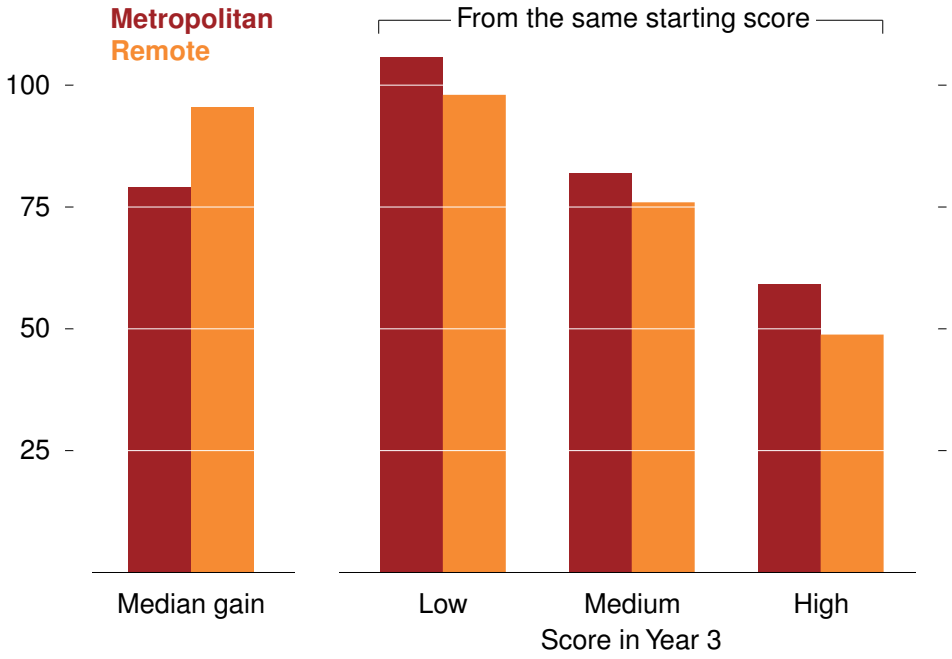


Notes: Similar patterns exist for other sub-groups, and at different percentiles. Gain scores estimated by a median quantile regression with cubic regression splines.
Source: Analysis of ACARA (2014).

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

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

Figure 3: Remote students make higher gains on average than metropolitan students, but lower gains from the same starting score
Median NAPLAN gain score between Year 3 (2012) and Year 5 (2014), reading, Australia



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

3 Looking at student progress through the lens of time

3.1 Standardised testing and grade equivalent scales

Student performance on standardised tests can be measured in a number of different ways.[11] The simplest measure, raw test scores, can be used to rank students. But raw scores can be hard to interpret. For example, on a 40-question test, the difference in skill level between a student with 25 correct answers and another with 20 correct answers should not be considered equal to the difference between a student with 40 correct answers and another with 35 correct answers. Raw test scores are even less useful for looking at student progress over time, because the measure does not take into account the degree of difficulty in the questions asked in different tests.

One alternative measure of student performance that used to be popular on standardised tests was the *grade equivalent scale*.[12] A test would be given to a sample of students across a range of schooling year levels (grades) to estimate the typical or expected

[11] Angoff (1984).
[12] Ibid.

score for students in each year level, usually the mean or median.[13] By interpolating across the estimates for each year level, the grade equivalent scale could also measure relative student performance in months. The appeal of such a measure is the intuitive interpretation – it is much clearer to say that a student is six months ahead of average than to say they are 20 points ahead of average, for instance.

A time-based measure similar to the grade equivalent scale is used on the *Programme for International Student Assessment* (PISA) – students are compared to the typical performance in terms of years and months of learning.[14] But the use of grade equivalent scales is not as common as it once was.[15] This is because there are a number of shortcomings with grade equivalent scales:

1. Individual student performance is measured with high variance (with standard errors often larger than one year of learning), particularly for high performers.
2. A number of students will score outside the range of tested outside the range of tested year levels; the grade equivalent scale must be extrapolated to estimate the relative performance of such students.
3. Grade equivalent scores are easily misinterpreted. For example, a Year 5 student who scores the same as a typical Year 8 student is estimated to be three years ahead of the average Year 5 student in terms of the skills tested. But this does not mean such a student would be comfortable with Year 8 material. In fact, we'd expect them to struggle since they haven't learnt the content in Years 6 and 7.

Scale scores based on the Rasch model, as used by NAPLAN and in many other standardised tests, do not have these shortcomings at the individual student level. It would be inappropriate to report student-level NAPLAN scores on a grade equivalent scale given these limitations. Yet just because one measure is more appropriate at one level – the individual student level – does not mean it is the most appropriate measure at every level. The issues with grade equivalent scales exist primarily at the individual student level. But policymakers are usually concerned with the performance and progress of particular groups of students rather than individual students. As the following section shows, a time-based comparison of student progress is more appropriate than a comparison based on the NAPLAN scale score.

[13] If the same test is given to each year level, there must be a large range of difficulty in the questions asked in order to get an accurate measure for high and low year levels.

[14] OECD (2013).

[15] We did find use of such a scale in an international reading test, see Renaissance Learning (2015).

3.2 Time-based comparisons lead to different conclusions

Consider two distinct groups of students: Group A and Group B. The scores displayed on Figure 4 on the next page are those of a representative student within each group (the median student): call these students A and B. Student A scores close to the average for numeracy, while Student B is below average, 103 NAPLAN points behind Student A in Year 3. Looked at in terms of NAPLAN points, as shown on the left chart, the gap between the students has reduced from 103 points in Year 3 to 71 points in Year 9.[16] At face-value, this suggests that Group B are catching up to Group A.

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

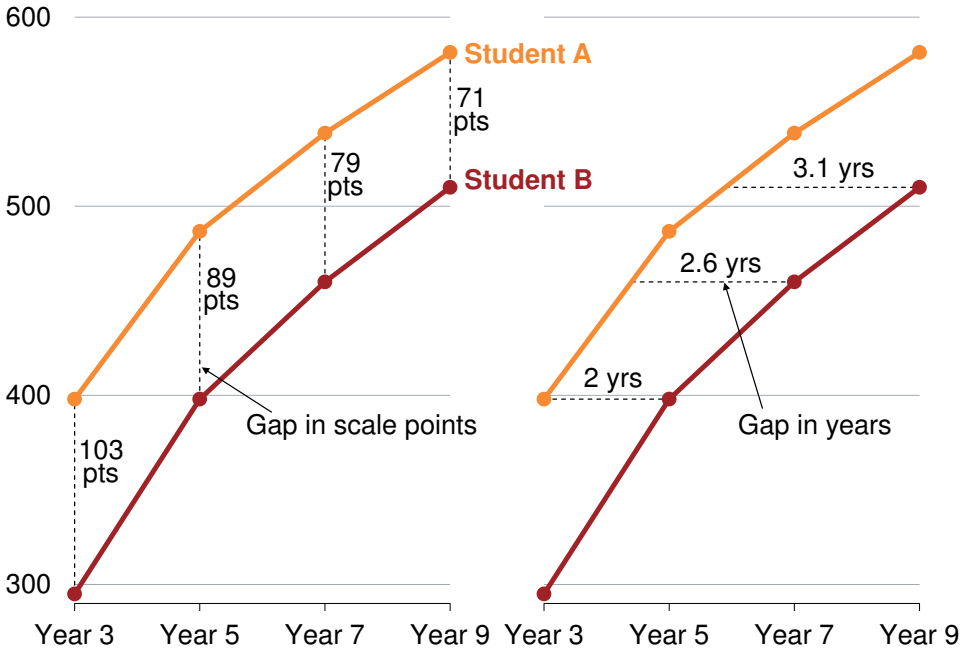
The difference between the students is defined in terms of Student A's rate of learning, but it could just as easily be defined in terms of Student B's rate of learning: "how long will it take Student B to reach the level of Student A?". While the story – that Student A learns comparable skills in less time than Student B – remains the same regardless of which student is defined as the benchmark, the size of the gap between the two in terms of 'years and months' is different. In Year 5, for instance, Student B is performing at Student A's level two years earlier, but Student B will take about three years to reach Student A's current level. We could say that Student A is two years ahead, but we could also say that Student B is three years behind.

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

[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 4: Measuring progress in years suggests a very different interpretation of NAPLAN results

NAPLAN scale score



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

3.3 Measuring *years of learning progress*

In order to construct a time-based measure of learning progress, it is first necessary to construct a grade equivalent scale. We use the median student in each year level as the benchmark to construct this measure with NAPLAN. It is easy to estimate scores corresponding to the test-taking years – Years 3, 5, 7, and 9. These are just the observed median scores in each test-taking year.[17] In Year 5 numeracy in 2014, for instance, the median NAPLAN scale score is approximately 489. A student with a score of 489 in any test-taking year is said to be performing at equivalent year level 5 (using 2014 as a reference year), meaning that their numeracy skill level is the same as a typical Year 5 student.

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

[17] Alternatively we could use the mean score, but the median score can be interpreted as the 'typical' student.

[18] This might not be the case for an individual student, but it is a reasonable assumption for the median of a large group of students.

This approach gives us a grade equivalent scale for Year 3 level up to Year 9 level. But if we are interested in estimating student progress for different groups from Year 3 to Year 9, this scale is not broad enough. Below-average groups will not have a grade equivalent level in Year 3, while above-average groups will have not have a grade equivalent level in Year 9. It is not possible for this scale to measure more than six years of progress between Year 3 and Year 9 for any student group. If we want meaningful comparisons, the grade equivalent scale must be extended beyond Years 3 and 9. But how can we estimate what a typical Year 10 student would score, for instance, without data on any Year 10 students?

The answer is to use linked student-level data. Some students in Year 7 are performing at a Year 9 level. We can make an assumption about such students – the median Year 7 student at a Year 9 level will make two years of progress between Years 7 and 9. This gives us an estimate of the NAPLAN scale score corresponding to Year 11 on the grade equivalent scale.[19] Similarly, some students in Year 5 are performing at a Year 4 level. If we assume the median Year 5 student at a Year 4 level made two years of progress between Years 3 and 5, then this gives us an estimate of Year 2 on the grade equivalent scale.

Using a median quantile 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). This gives us a broader grade equivalent scale that can be used the measure the years of learning progress between Years 3 and 9 for many student groups.[20] It is possible to extrapolate the curve further than two years ahead of Year 9, but the results are less robust at such points.

Figure 5 on the following page shows conceptually how these approaches are used to construct a curve that maps NAPLAN scale scores to estimated equivalent year levels. The methodology is described in detail in Appendix B on page 23, and the accuracy of the assumptions are explored further.

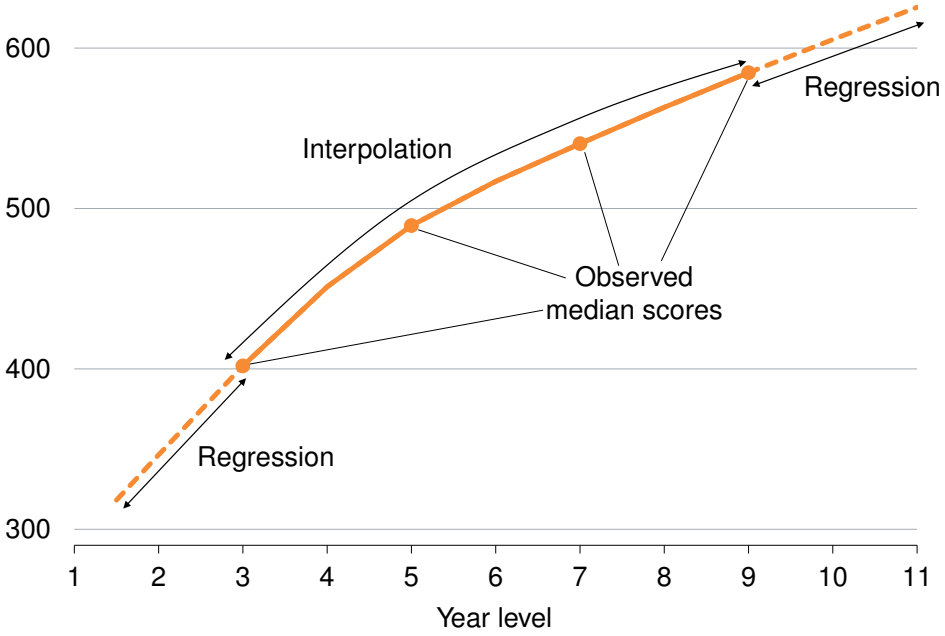
Having constructed the benchmark curve, it is possible to track the equivalent years of progress made by a given student or a group of students. An example of this is shown in Figure 6 on the following page for the median student (Student A) of an above-average student group. In Year 3, this student is about one year and seven months ahead of the benchmark curve in Year 3. By tracking Student A back to the benchmark curve,

[19] This is not necessarily the score we'd expect from a typical Year 11 student, since the curriculum changes significantly at Year 11. But it is reasonable to say that this score is two years above Year 9 level.

[20] When students are divided into sub-groups based on family or school background, the median student in each group is normally ahead of equivalent year level 1.5 in Year 3, and below equivalent year level 11 in Year 9. Hence, this scale is broad enough to compare student progress for most of these groups.

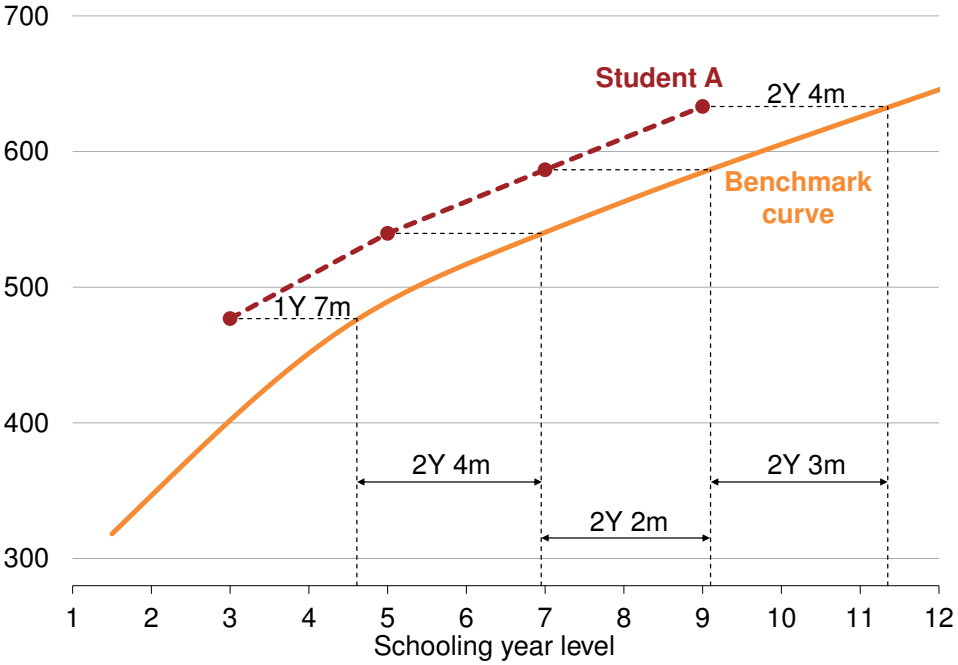
we can conclude that this group made above-average progress between each NAPLAN test, finishing Year 9 two years and four months ahead of the benchmark. That is, this student made six years and nine months of progress between Year 3 and Year 9.

Figure 5: Estimating the equivalent year level benchmark curve involves interpolation and regression
Estimated median NAPLAN scale score, numeracy, Australia



Source: Analysis of ACARA (2014).

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



Source: Analysis of ACARA (2014).

4 Results

An analysis using *years of learning progress* is presented in *Widening gaps: what NAPLAN tells us about student progress*.[21]

5 Extensions

It would be possible to construct multiple benchmark curves to track the progress of students or schools relative to a benchmark student with similar characteristics. This is an area of further research.

[21] Goss et al. (2016). Report available at www.grattan.edu.au/widening-gaps.

A Appendix: Data sources and issues

A.1 Student-level NAPLAN datasets

There are two major datasets used in the analysis:

- NAPLAN results across all four assessment domains and year levels for all Australian students recorded in 2014, linked with their 2012 results where applicable.[22] This dataset contains test scores for more than one million students for each domain in 2014, and more than 700,000 in 2012.[23]
- NAPLAN results across all four domains recorded between 2009 to 2015 for the cohort of Victorian students who were in Year 3 in 2009.[24] 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.

The grade equivalent scale is estimated using the national dataset – this creates a national benchmark for student progress. We then use this benchmark to analyse student progress in 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.[25]

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).[26] The national dataset contains a randomised school-level indicator, but it is not possible to identify schools themselves.

[22] ACARA (2014).

[23] 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.

[24] VCAA (2015).

[25] This allows the analysis to pick up Victorian-specific effects. It should be noted that, on average, Victorian students score higher than most other states. One explanation for this is that Victorian students are, on average, more likely to come from a high SES background [ACARA (2014)].

[26] To prevent school identification, the Victorian ICSEA data were given to us in bands of 26 points.

A.2 Defining the 'typical' student

The analysis focuses on the 'typical' student, either at the population level or within a particular sub-group of students, defined as the student with the median NAPLAN scale score. Analysis of particular sub-groups of students (such as those grouped by parental education) is performed according to the typical student within each sub-group – the sub-group median.

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

A.3 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.

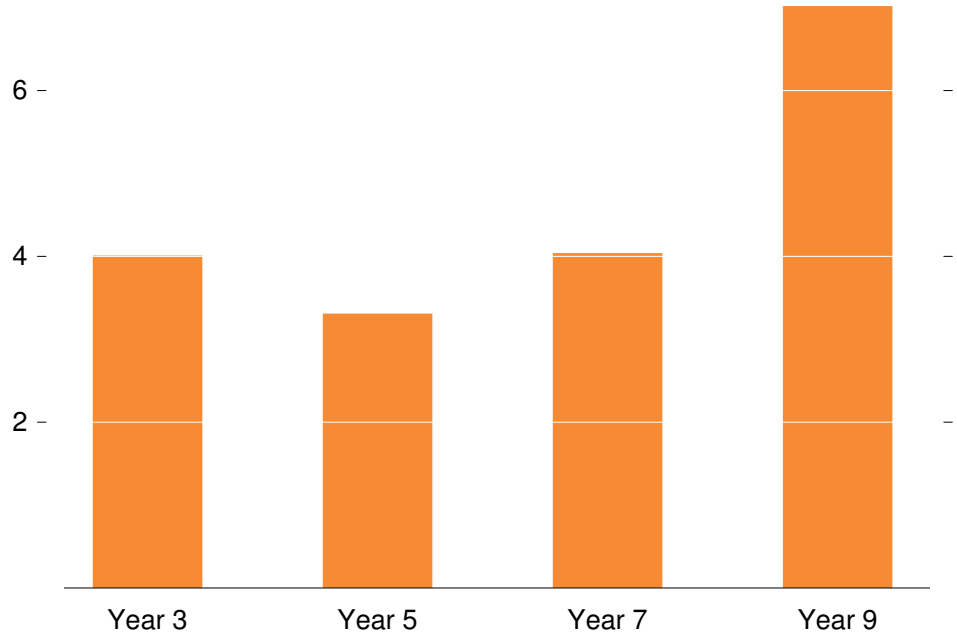
For any given NAPLAN test, participation rates are high, usually exceeding 90 per cent. The most common reason for non-participation is student absenteeism. This is usually four per cent or less, but rises to seven per cent in Year 9, as shown for numeracy in Figure 7. A small proportion of students (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, and also because some students moved out of Victoria before Year 9.[28]

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

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

Figure 7: 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: Analysis of VCAA (2015).

A brief analysis suggests that students are less likely to miss a test due to being absent/withdrawn or an exemption if their parents are better educated. Figure 8 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 one test between Year 3 and Year 9, compared to only 25 per cent of students where a parent has a university degree.

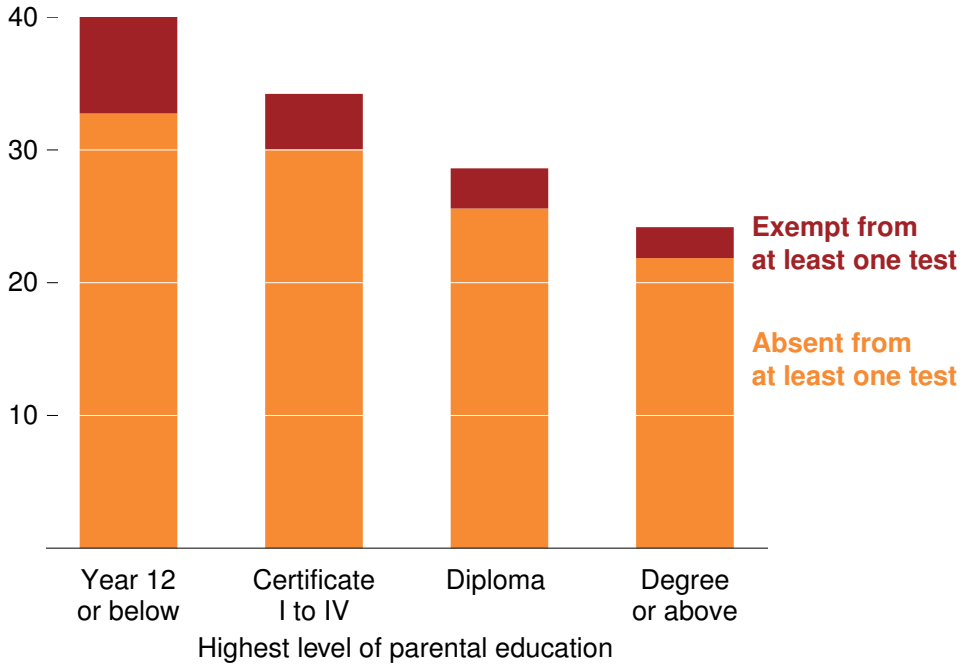
Given that students of well-educated parents typically score higher and make higher gains from a given starting score than those whose parents are less well educated, the consequence of ignoring missing data is an upwards bias in estimates of the median score and median gain score.[29]

It is also possible that students who miss a test would have made a lower gain score than other students, even after controlling for starting score. With only two years of linked data it would not be possible to test this. But with four years of linked data, as is available with the Victorian 2009 to 2015 cohort, there are students that have missed a test in one or two years, but for whom we observe NAPLAN scale scores in at least two other years. Figure 9 on page 18 shows the estimated median gain score in reading between Years 5 and 7 for students that did not miss a test in any year, and for students that missed a test in Year 3, Year 9 or both. Not only are those that

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

Figure 8: Students from households with higher parental education are less likely to miss one or more NAPLAN tests

Percentage of students that miss a NAPLAN test, Victorian 2009–15 cohort



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

missed a test predicted to make smaller gains, but the gap is larger for students whose parents do not have a degree or diploma.

This means that estimates of median progress for particular sub-groups are likely to be upwards biased if missing data are ignored. But the bias is likely to be much larger for lower levels of parental education. In turn, this means the gap in student progress calculated between students with high and low parental education is likely to be underestimated rather than overestimated.[30]

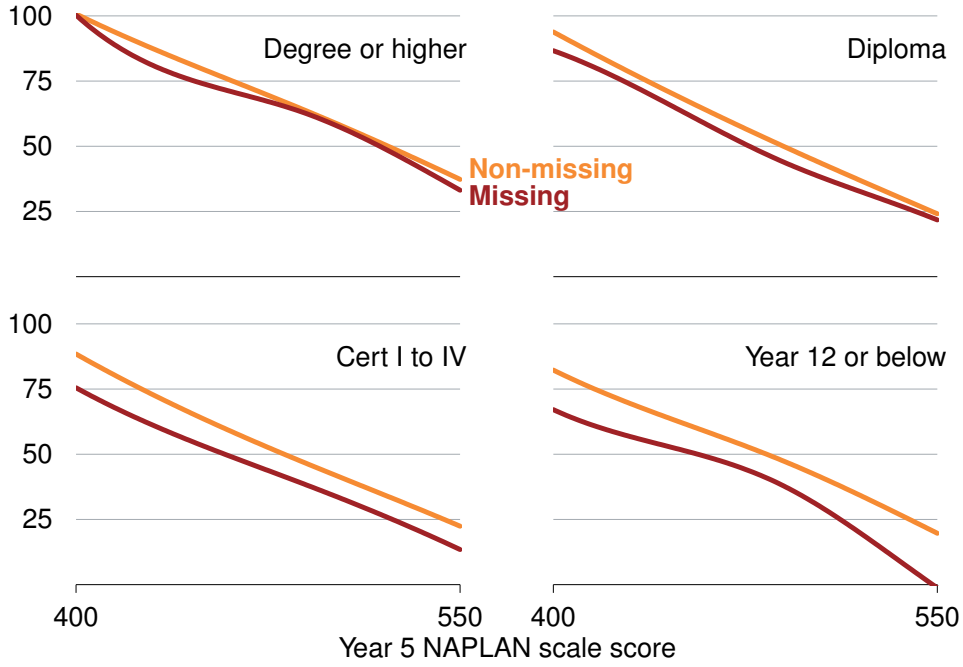
Our analysis of NAPLAN gain scores does not impute missing results. Students who are given an exemption from one or more tests are excluded from the analysis.[31]

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

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

Figure 9: Missing data have more of an impact on gain scores for students from less-educated households

Median NAPLAN gain score by highest level of parental education, reading, Year 5 to Year 7, Victorian 2009–15 cohort



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

A.4 Measurement error and bias

A.4.1 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).[32] 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.[33] 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.

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

[33] 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 (2015c)].

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.[34]

A.4.2 Using NAPLAN scale scores (WLEs) may result in imprecise estimates of progress

Skill level is continuous, but NAPLAN scale scores are discrete

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

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

Regression to the mean

In the context of comparing student progress over two or more NAPLAN tests, *regression to the mean* suggests that an extreme NAPLAN score in one year (either extremely low or high) is likely to be followed by a less extreme score on the following test (two years later). This is not because students at the extremes are making significantly high or low progress, but because the original test score is exaggerated by measurement error. This may lead to learning progress being significantly overstated by gain scores for students who start with a very low score, and understated for students who start with a very high score.[36]

[34] 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.

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

[36] 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 2 on page 6 and Figure 9 on page 18. But if this

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.[37]

It is important to note that an extreme score for a particular sub-group might not be an extreme score for another sub-group. For example, the NAPLAN scale score equal to the 95th percentile in Year 7 numeracy for those whose parents have no post-school qualifications is only at the 82nd percentile for those who have a parent with a university degree. This means that the regression to the mean between the Year 7 and Year 9 test is likely to be stronger for a high achieving student whose parents have no post-school qualifications than it is for a high achieving student with a university-educated parent.[38]

A.4.3 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:

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

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.[39] This creates a sample that has the properties of a continuous distribution, allowing for more accurate estimates of percentiles.

---

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.

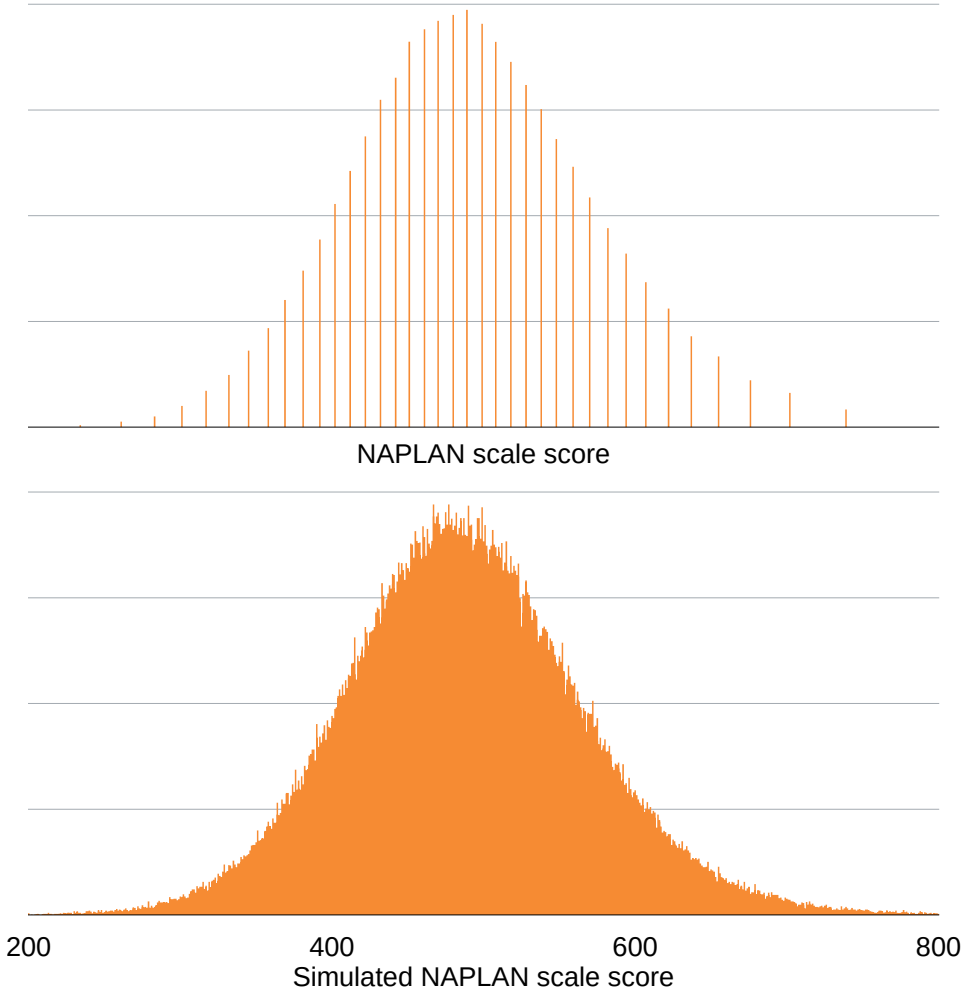
[37] A way to think about this is that the effective number of questions declines as student skill level moves further from the level at which the test is set. For example, a student at the 90th percentile will find most questions too easy, while a student at the 10th percentile will find most questions 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.

[38] Just because a student does not have a university-educated parent, this does not mean that a high NAPLAN scale score is overstating their true skill level. But when we compare two students with the same high score, one with a university-educated parent and one without, the one without is more likely to have had an unusually good test day (i.e. scoring above their true skill level) than the student with a university-educated parent.

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

Figure 10 compares a histogram of discrete NAPLAN scale scores to a histogram of simulated NAPLAN scale scores. While this approach does not remove measurement error at the individual student level, it takes into account that measurement error varies across students with different scores.

Figure 10: The simulation approach solves the issue of discrete NAPLAN scale scores
Histogram of Year 5 NAPLAN scale score, numeracy, Australia



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

Avoiding extreme percentiles

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

given domain is between 0.75 and 0.8 – this strong correlation suggests regression to the mean will have only a small impact for most percentiles.

Nonetheless, our analysis aims to avoid estimating NAPLAN scale scores and gain scores for students at extreme percentiles, and most analysis is focused around the median student. We use a rule of thumb to minimise bias due to regression to the mean – no analysis is based on the estimated NAPLAN scale score or gain score of students below the 10th percentile or above the 90th percentile.[40]

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

Calculating standard errors

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.[41] 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.[42]

Plausible values

The best approach to reduce the impact of measurement error is to use *plausible values*. Like the simulation approach outlined above, this approach would simulate a NAPLAN scale score from a continuous distribution for each student, including imputing values for missing data. But plausible values are simulated from a distribution that takes into

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

[41] 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.

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

account student and school background factors.[43] NAPLAN reports produced by ACARA are based on analysis using plausible values.[44]

When simulated correctly, plausible values are able to produce unbiased estimates of percentiles and gain scores for each sub-group.[45] Plausible values were available for the 2014 test year in the national dataset, but not for the 2012 results or the Victorian 2009–15 cohort. This means we did not have the data to use plausible values to analyse progress.[46]

B Appendix: Mapping NAPLAN scale scores to a grade equivalent scale

B.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.[47] 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.

A new framework to interpret NAPLAN results is presented in this paper. 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.[48]

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.

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

[44] ACARA (2015d), p. 22.

[45] Wu (2005).

[46] 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.

[47] 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.

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

B.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}).[49]

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

$$\begin{aligned} f_j(Y) &= Q_{50}\left[X_j \mid Y\right] \\ y_1 < y_2 &\implies f_j(y_1) < f_j(y_2) \\ f_j(Y) &\in [f_j(y_{\min}), f_j(y_{\max})] \end{aligned} \tag{2}$$

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.[50]

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

$$Y^* = f_j^{-1}\left(X_j\right) \tag{3}$$

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

B.3 Estimating equivalent year levels

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

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

[50] 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.

these year levels.[51] But with the tests only administered in four year levels, we must make further assumptions to estimate $f_j(Y)$.

The methodology estimates $f_j(Y)$ (the median NAPLAN scale scores corresponding to a given year level) using the simulated NAPLAN results (see Section A.4.3) 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 at year levels 3, 5, 7, and 9

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

$$\begin{aligned} \widehat{f}_j(3) &= \tilde{x}_{j,3} \\ \widehat{f}_j(5) &= \tilde{x}_{j,5} \\ \widehat{f}_j(7) &= \tilde{x}_{j,7} \\ \widehat{f}_j(9) &= \tilde{x}_{j,9} \end{aligned} \tag{4}$$

where $\tilde{x}_{j,y}$ is the sample median NAPLAN scale score in year level y.[52]

Step 2: Interpolate between Year 3 and Year 9

Using a third-order polynomial, fit a smooth curve through the four data points, $([Y, \widehat{f}_j(Y)], Y = 3, 5, 7, 9)$, to estimate $f_j(Y)$ between Year 3 and Year 9, as shown in Figure 11.

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

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

$$g_{j,Y}\left(X_{j,Y-2}\right) = Q_{50}\left[X_{j,Y} - X_{j,Y-2} | Y, X_{j,Y-2}\right] \tag{5}$$

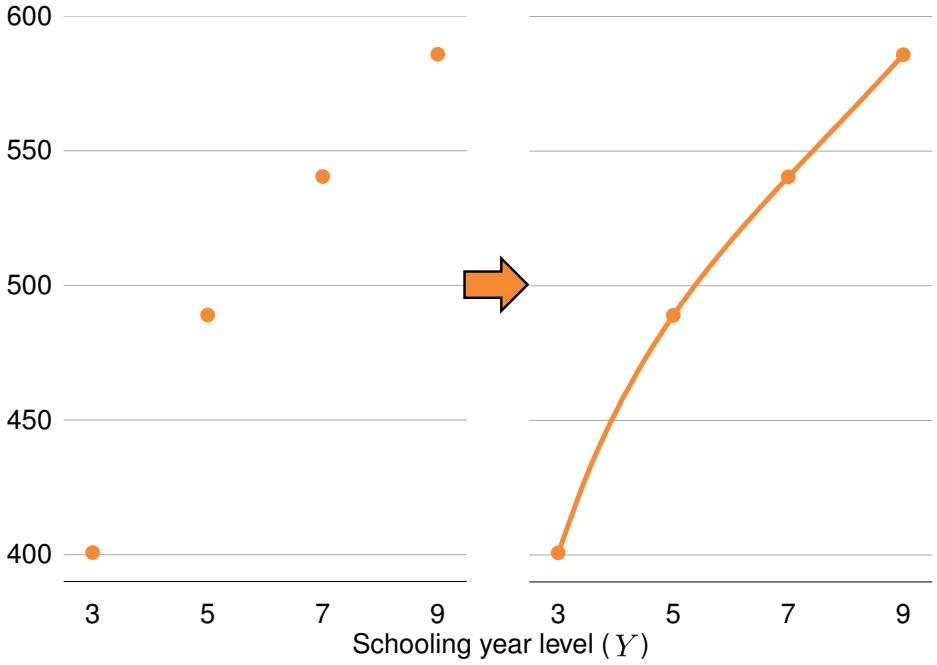
where $X_{j,Y}$ denotes NAPLAN scale score in domain j in school year Y. For students that scored $x_{j,3}$ in Year 3 reading, for example, $g_{j,5}(x_{j,3})$ is the median gain score these students will make to Year 5.[53]

[51] This is a useful way of thinking about what equivalent year levels are trying to measure. But it is important to note that the interpretation of equivalent year levels 11 and 12 estimated with the available data could be very different to those estimated with data on Year 11 and Year 12 students.

[52] 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.

[53] 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 11: A third-order polynomial is used to interpolate between Year 3 and Year 9
Estimated median NAPLAN scale score, $\widehat{f}_j(Y)$, numeracy, Australia



Source: Analysis of ACARA (2014).

From eqs. (2) and (5), it follows that:

$$g_{j,Y}\left[f_j(Y-2)\right] = f_j(Y) - f_j(Y-2) \tag{6}$$

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

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

$$\begin{aligned} g_{j,Y}(X_{j,Y-2}) = \beta_0 + \beta_1 X_{j,Y-2} + \beta_2 S_2(X_{j,Y-2}) \\ +\beta_3 S_3(X_{j,Y-2}) + \beta_4 S_4(X_{j,Y-2}) \end{aligned} \tag{7}$$

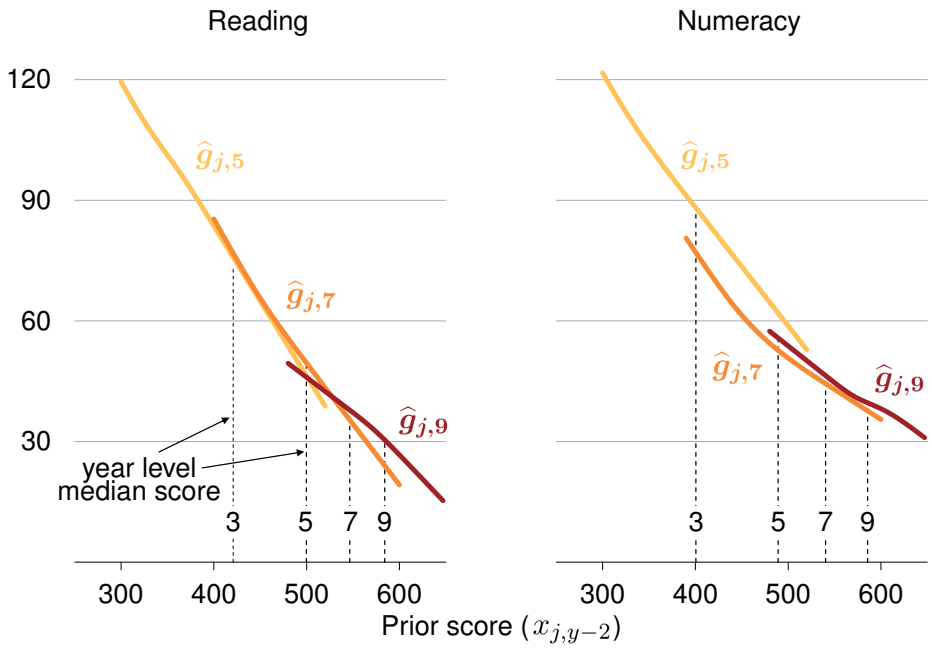
where S_2, S_3 and S_4 are functions that create spline variables.[54] Alternatively, this function could be specified with quadratic or higher order polynomial terms.

Given $g_{j,Y}$ represents a conditional median gain score, eq. (7) can be thought of as a quantile regression model at the median. This can be estimated using least absolute deviations.[55]

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

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

Figure 12: The estimated median gain score is strongly related to prior score, but only weakly related to year level
Two-year median NAPLAN gain score, $\widehat{g}_{j,y}(x_{j,y-2})$, Australia



Source: Analysis of ACARA (2014).

Figure 12 plots the estimated functions, $\widehat{g}_{j,y}\left(x_{j,y-2}\right)$, for $y = 5, 7$ and 9 for both reading and numeracy. Predicted median NAPLAN gain scores are much higher for lower prior scores, but year level does not have a large effect on gain scores once prior scores are controlled for. For instance, when evaluated at the NAPLAN 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. (6) gives:

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

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

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

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

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

$$g_{j,11}\left[f_j\,(9)\right] \approx g_{j,9}\left[f_j\,(9)\right] \tag{10}$$

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:

$$g_{j,4}\left[f_j\,(2)\right] \approx g_{j,5}\left[f_j\,(2)\right] \tag{11}$$

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

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

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

$$\begin{aligned} \widehat{f}_j\,(10) &= \widehat{f}_j\,(8) + \widehat{g}_{j,9}\left[\widehat{f}_j\,(8)\right] \\ \widehat{f}_j\,(11) &= \widehat{f}_j\,(9) + \widehat{g}_{j,9}\left[\widehat{f}_j\,(9)\right] \end{aligned} \tag{12}$$

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. (11) and its extensions, $f_j(1.5), f_j(2)$ and $f_j(2.5)$ are estimated by solving the following equations for $\widehat{f}_j\,(Y)$:

$$\begin{aligned} \widehat{f}_j\,(1.5) &= \widehat{f}_j\,(3.5) - \widehat{g}_{j,5}\left[\widehat{f}_j\,(1.5)\right] \\ \widehat{f}_j\,(2) &= \widehat{f}_j\,(4) - \widehat{g}_{j,5}\left[\widehat{f}_j\,(2)\right] \\ \widehat{f}_j\,(2.5) &= \widehat{f}_j\,(4.5) - \widehat{g}_{j,5}\left[\widehat{f}_j\,(2.5)\right] \end{aligned} \tag{13}$$

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.[56] 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.[57]

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

$$\widehat{Y}^* = \widehat{f}_j^{-1}(X_j) \tag{14}$$

Figure 13 shows this curve for reading and numeracy, both in terms of $\widehat{f}_j(Y)$ and in terms of its inverse, $\widehat{f}_j^{-1}(X_j)$. As the chart on the right shows, every NAPLAN score (within the range of the curve) can be mapped to an equivalent year level. A score of 500 in numeracy, for instance, corresponds to an equivalent year level of 5 years and 4 months – a student at this level can be interpreted as performing four months ahead of the typical (median) Year 5 student at the time of the Year 5 NAPLAN test.[58]

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

B.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:

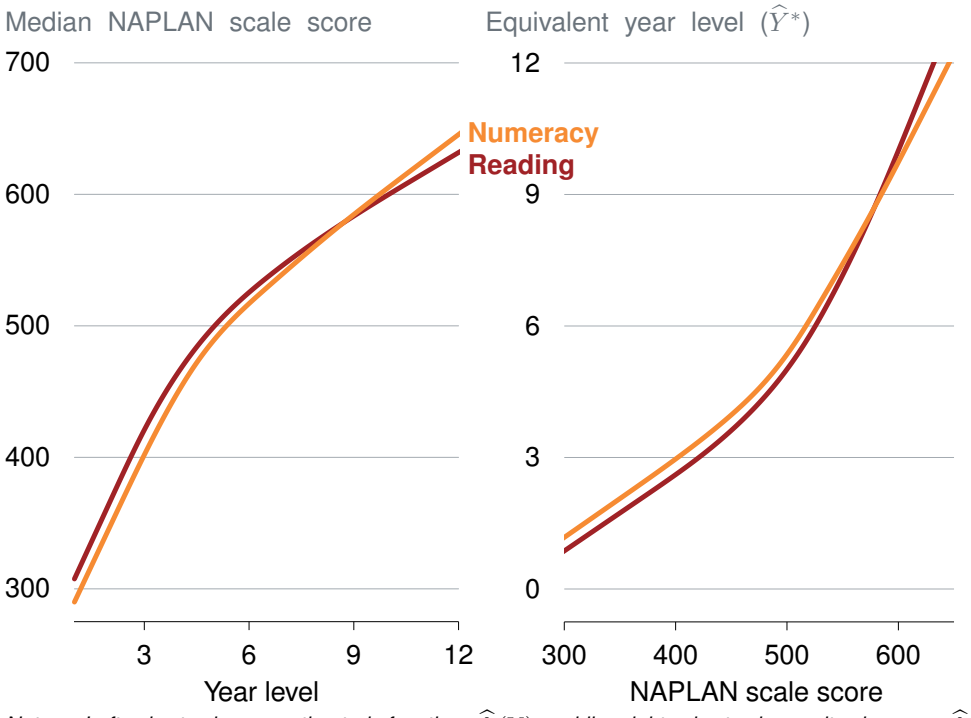
- what is the standard error at different points along the equivalent year level curve?

[56] 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.

[57] Given the estimated curve, $\widehat{f}_j(Y)$ is approximately concave between Year 1.5 and Year 11, we would expect concavity to hold if the curve is extended to Year 1 and Year 13. As such, linear extrapolation is unlikely to underestimate the median scale score for Year 1, Year 12, and Year 13 – this is conservative for estimating the gaps in progress between different groups.

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

Figure 13: All NAPLAN scale scores in a given range correspond to an equivalent year level



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

- 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.

B.4.1 Standard errors around point estimates

There are two sources of error that the standard error accounts for: test measurement error for individuals, and the error associated with a finite sample. But the equivalent year level curve is calculated from a very large sample, meaning that the standard error around estimates of the median is naturally small.[59]

In reporting, we prefer using confidence intervals to standard errors, since equivalent year levels are asymmetrically distributed around NAPLAN scale scores. We calculate

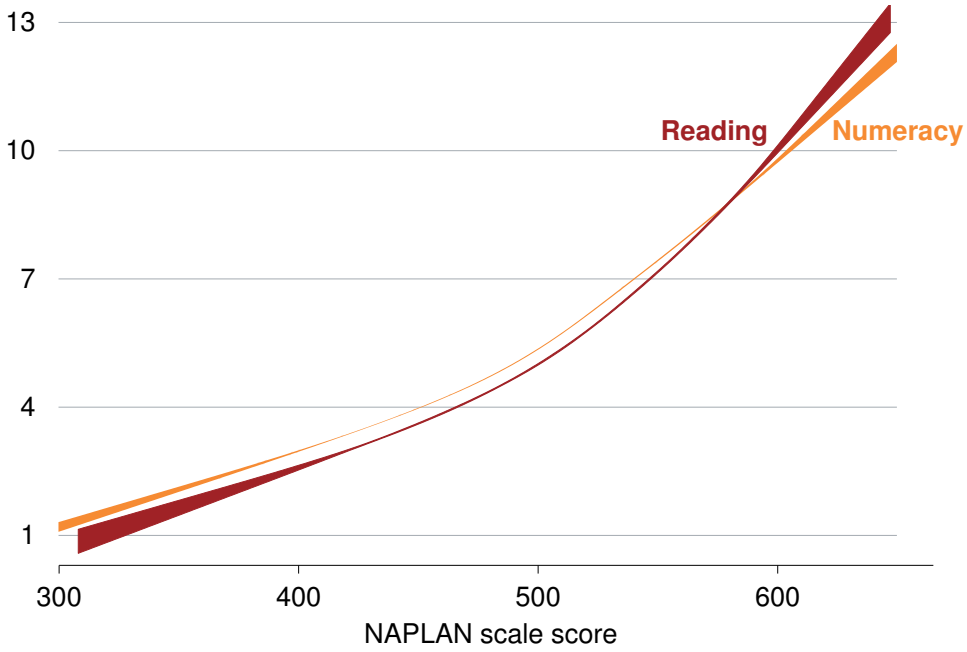
[59] This assumes that individual measurement error is not systematically biased.

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.[60]

Between Year 3 and Year 9, equivalent year levels are estimated with a very narrow confidence interval. As the curve is flatter in Year 9 than it is in Year 3, the confidence interval around Year 9 is wider. The width of the confidence interval naturally increases below Year 3 or above Year 9. For a score of just over 300 in reading (close to equivalent year level 1), the 99 per cent confidence interval around the equivalent year level estimate is about seven months of learning, while for a score of 650 (close to equivalent year level 13), the 99 per cent onfidence interval is eight months.[61] But for scores between 400 and 600, the 99 per cent confidence interval does not exceed two months of learning. These intervals are displayed in Figure 14 and table 1.

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

Figure 14: Confidence intervals are much wider in the extremes
Estimated 99 per cent confidence interval for equivalent year levels, Australia



Source: Analysis of ACARA (2014).

[60] 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.

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

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

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

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

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

Without students taking a NAPLAN test outside of the test-taking years, it is impossible to validate whether our estimates of the median NAPLAN scale score in Years 2, 10, and 11, for instance, reflect how the median student would actually perform in those year levels.[62] But it is possible to use a similar methodology to predict the median score in Year 3 and Year 9 without using data from Year 3 and Year 9. This can then be compared to the estimated median NAPLAN scale score for Year 3 and Year 9 on the full dataset.

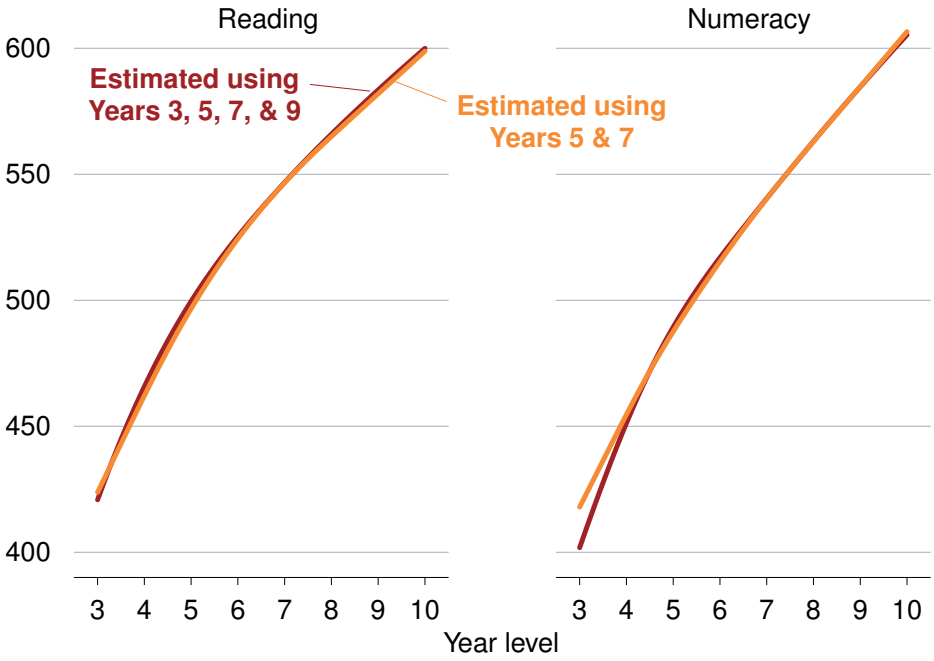
Using data for students in Year 7 linked to their Year 5 results, Figure 15 shows that the methodology predicts the median NAPLAN scale score outside these year levels with reasonable accuracy (using the curve based on the full dataset as a benchmark). There is some evidence, however, that predicting the median score for year levels well beyond the available data will lead to inaccuracies.[63]

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

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

Figure 15: Data from Years 5 and 7 students provides a reasonable approximation for other year levels

Estimated median NAPLAN scale score, Australia



Source: Analysis of ACARA (2014).

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

B.4.3 How do estimates change with different assumptions?

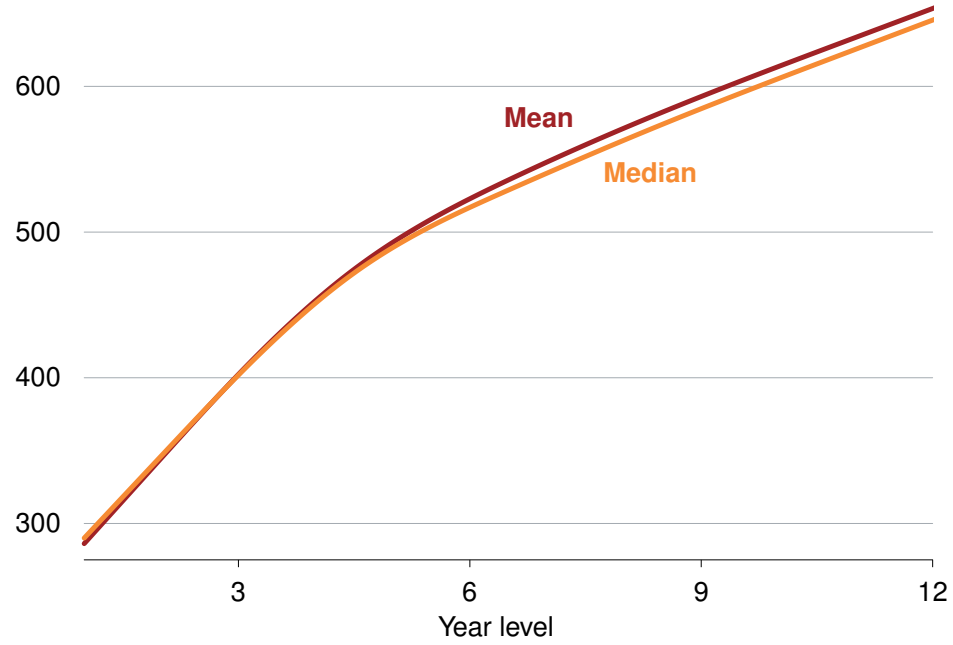
Using a different benchmark student

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

The most obvious alternative to using the median is to use the mean NAPLAN scale score in each year level. This has a noticeable, but relatively small impact on the shape of the curve, as shown in Figure 16.[64]

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

Figure 16: Using the mean instead of the median changes the curve slightly
Estimated median NAPLAN scale score, numeracy, Australia



Source: Analysis of ACARA (2014).

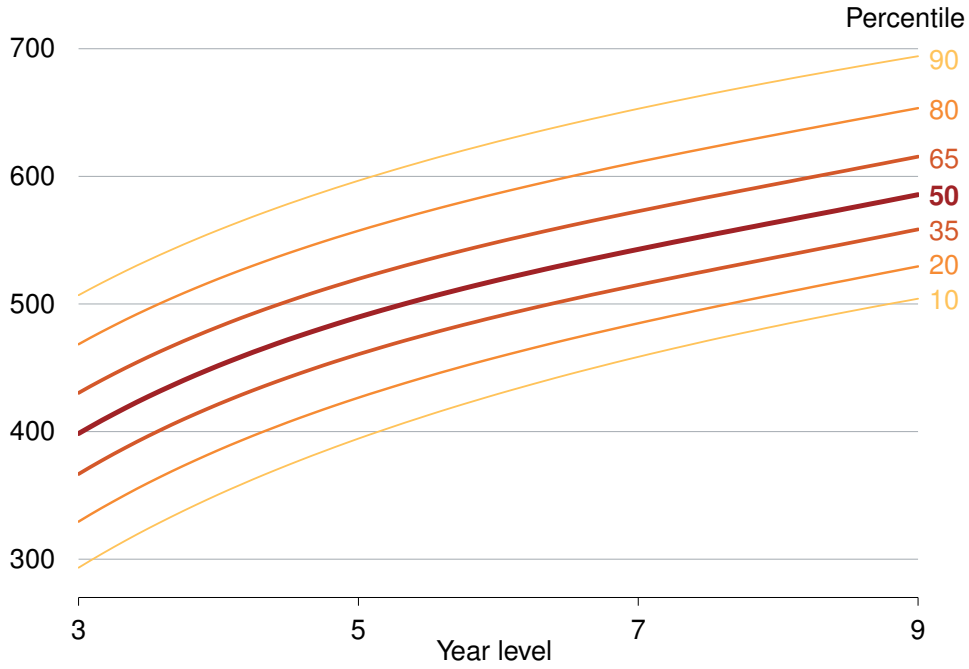
Alternatively, instead of using a measure of central tendency such as the mean or median, the benchmark could be set much higher – say, at the 80th percentile. A *year of progress* would then be something harder for students to attain, but could be seen as something to aspire to. A curve based on the 80th percentile would be a better way of grouping high achieving students (for instance, those with NAPLAN scale scores above 650 in Year 9), but it would be difficult to accurately estimate what the 80th percentile student would have scored on a NAPLAN test taken before Year 3. Thus, this curve is unlikely to provide a good measure of progress over six years for average and below-average students. In any case, it is worth noting that all percentiles between the 10th and the 90th appear to be concave, as shown in Figure 17 on the next page.

Using control variables to estimate gain scores

One assumption that was strongly considered in this methodology was to include control variables in eq. (7) – 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 benchmark curve that is steeper for lower scores, and flatter for higher scores. While using control variables makes intuitive sense, when $g_{j,Y}$ is

Figure 17: All percentiles make smaller gain scores at higher year levels
NAPLAN scale score by percentile, numeracy, Australia



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

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.[65]

Treatment of missing data

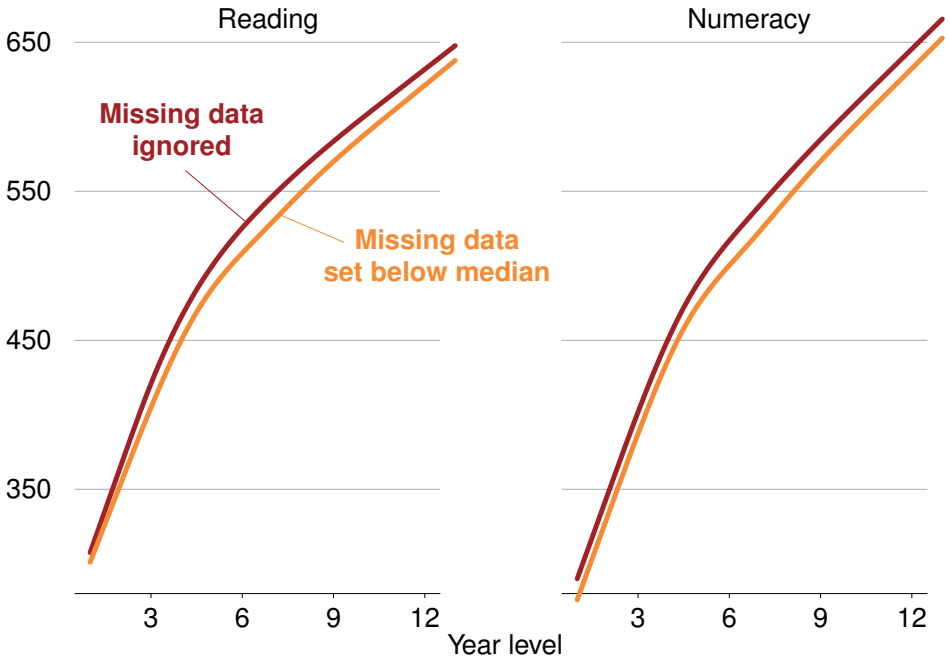
Students that are exempt, absent, or withdrawn from a NAPLAN test in either 2012 or 2014 are ignored for the purposes of estimating the median NAPLAN scale score in each year level. But Section A.3 suggests that students who miss a test are more likely to come from households with lower parental education, and are likely to make smaller gain scores from a given prior score than other students. This means the estimated median score is likely to be above the true 50th percentile.

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

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

Figure 18: Treating missing data as below the median does not change the shape of the curve

Estimated median NAPLAN scale score, Australia



Source: Analysis of ACARA (2014).

Figure 18 shows that this alternative treatment of missing data will, unsurprisingly, lead to a lower estimate of the median NAPLAN scale score in each year level. But the curves for both reading and numeracy still have the same concave shape. It is unlikely that this alternative treatment of missing data would lead to very different conclusions about the gaps in student progress.

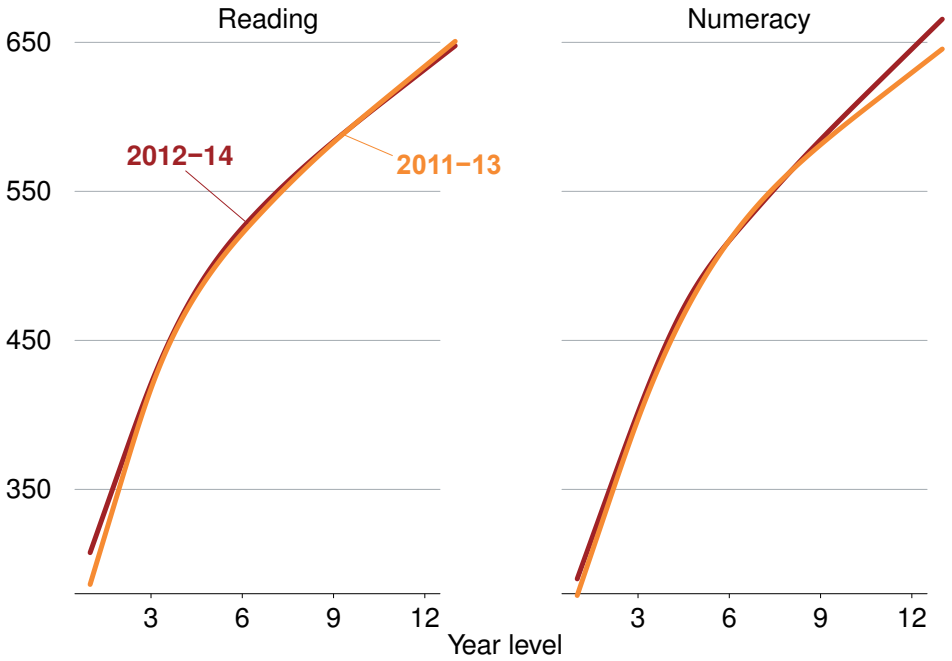
B.4.4 How robust are estimates to different cohorts

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.[66] At the national level, results are not expected to change significantly across two cohorts one year apart.

We cross-checked our results by applying the methodology to the national cohort of 2013 students, with results linked to 2011. As Figure 19 shows, in reading, the 2011-13 results are almost identical to those of 2012-14, except for Year 1 where the standard error is high (see Section B.4.1). In numeracy, there is little noticeable difference below Year 9, but the estimated curve using the 2011-13 data is flatter for later year levels. This means the 2012-14 numeracy curve will provide more conservative estimates of

[66] 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 19: There are some discrepancies that arise with different cohorts
Estimated median NAPLAN scale score, Australia



Source: Analysis of ACARA (2013) and ACARA (2014).

progress for high achievers, students with high levels of parental education and students from high advantaged schools.

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

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

Of course, it is also important to consider the limitations of this approach. In terms of the methodology outlined in this chapter, equivalent year levels are not an appropriate way of reporting individual student results. This is because equivalent year levels do not cover the full range of NAPLAN scale scores, so this measure is inappropriate for high-achieving students (those performing above equivalent year level 12). In addition, high levels of measurement error at the individual level mean that it is difficult to accurately assign a student to an equivalent year level.[67]

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

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

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

References

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). *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.

— (2015b). *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.

— (2015c). *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.

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

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

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

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.

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

VCAA (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.