Technical Report 2019: Approach to reporting on My School

My School®

2020

Contents

1. Introduction

In June 2019, the Council of Australian Governments (COAG) Education Council considered the NAPLAN Reporting Review prepared by Emeritus Professor Bill Louden and tasked ACARA with a series of goals relating to technical aspects of NAPLAN and the provision of information via the *My School* website. Following consultation with the National Assessment, Data, Analysis and Reporting Reference (NADAR) Group and Data Strategy Group (DSG), a series of recommendations were proposed:

- The utilisation of ICSEA in a new way to improve the explanation of differences in school performance, without substantial increases in the amount of data collection.
- A model based on ICSEA data be used to create a predicted score for each school's NAPLAN results based on the results of all students with a similar socio-educational background. This predicted score, rather than the average result from only 60 schools, would then become the point of comparison with a school's actual results for the purposes of *My School* benchmarking.
- The development of an additional gain measure that would provide an alternative perspective on the school's performance in improving student outcomes over the last two years. This would benchmark the gain achieved by students at the selected school against the average gain achieved by students who had both the same starting score and similar socio-educational advantage. This measure would assist alleviate the criticism that a simple gain measure is not adequate to address the pattern that learning trajectories are different for students with different starting points.

The following report outlines the methodological changes undertaken to address these recommendations.

2. Socio-educational advantage (SEA)

For the majority of Australian students, data were collected about their parents' level of education and occupation. It is well known that parental education and occupation have a strong, positive relationship with student academic achievement.

The national student background data (SBD) collection included four variables indicating the educational and occupational background of each parent. The four variables and their respective values were:

1. School education
    a. Year 9 or less
    b. Year 10
    c. Year 11
    d. Year 12
2. Non-school education
    a. No non-school education
    b. Certificate I-IV, incl. trade certificate
    c. Advanced diploma / Diploma
    d. Bachelor's degree or above
3. Occupation group
    a. Machine operator
    b. Trades person / Clerk / Sales
    c. Professional / Manager
    d. Senior Manager
4. Non-paid work
    a. Paid work
    b. Non-paid work

Many students' parents reported data on all eight variables: four for each parent. However, the data often included empty fields for some or all variables. In 2013, a methodology was developed by ACARA to combine the eight variables into one index—socio-educational advantage (SEA)—and to more effectively address partially or fully missing parental data than previous methodologies. The generalised partial credit model from the broader Item Response Theory (IRT) framework was used as implemented in ACER ConQuest V4. The methodology is described in detail in Section 3 of the ICSEA 2013: Technical report.

For the newly developed *similar-students analysis* and *progress rate analysis* described in this report, three adjustments were made to the 2013 SEA model.

Firstly, in the 2013 SEA model, data were not weighted when the parental occupation and education items were calibrated. Without these weights, the scale calibration was dominated by larger jurisdictions. To be consistent with similar statistical models used in NAPLAN and NAP sample studies, and to address concerns from smaller jurisdictions regarding the validity of the model for their context, a senate weight was used. By applying these senate weights all jurisdictions contributed equally to the scaling process.

Secondly, some unlikely combinations of occupation codes with other variables were recoded to missing in the ACT. It was observed that the correlation between non-school education and occupation groups was lower in the ACT and the Northern Territory compared to other jurisdictions. Given that the combination was unlikely, but not impossible, it was decided not to apply a blind recode for all jurisdictions. There were two reason for applying the recode to ACT only. The first reason was that relatively more parents with lower levels of education had selected the highest occupation group in the ACT than in other jurisdictions. The second reason was that, when comparing distributions of parental education and occupation between NAPLAN data and ABS data, it became apparent that the percentage of senior managers in the ACT was higher in NAPLAN data than in ABS data. No evidence was found for either of these anomalies in any of the other jurisdictions. While these issues require further investigation, it was decided by ACARA that very unlikely occupation codes should be removed from the ACT data prior to the 2019 analysis. That is, the senior manager occupation group was set to missing for parents who indicated their highest level of schooling was Year 10 or below, or who had not completed any non-school education. This resulted in recoding of two per cent of parental occupation codes in the ACT.

Thirdly, information was added to the model indicating which students were from the Northern Territory. The amount of missing data was large in the NT relative to other jurisdictions and SEA scores were relatively low. Therefore, the addition of a Northern Territory indicator to the model improved imputation of SEA values for students from the Northern Territory with missing parental data. While it is beyond the scope of this technical report to describe in detail the theoretical foundations of plausible values methodology, some explanation about the inclusion of this regressor is included in Appendix A, together with more technical details of the adjustments made to the 2013 SEA model.

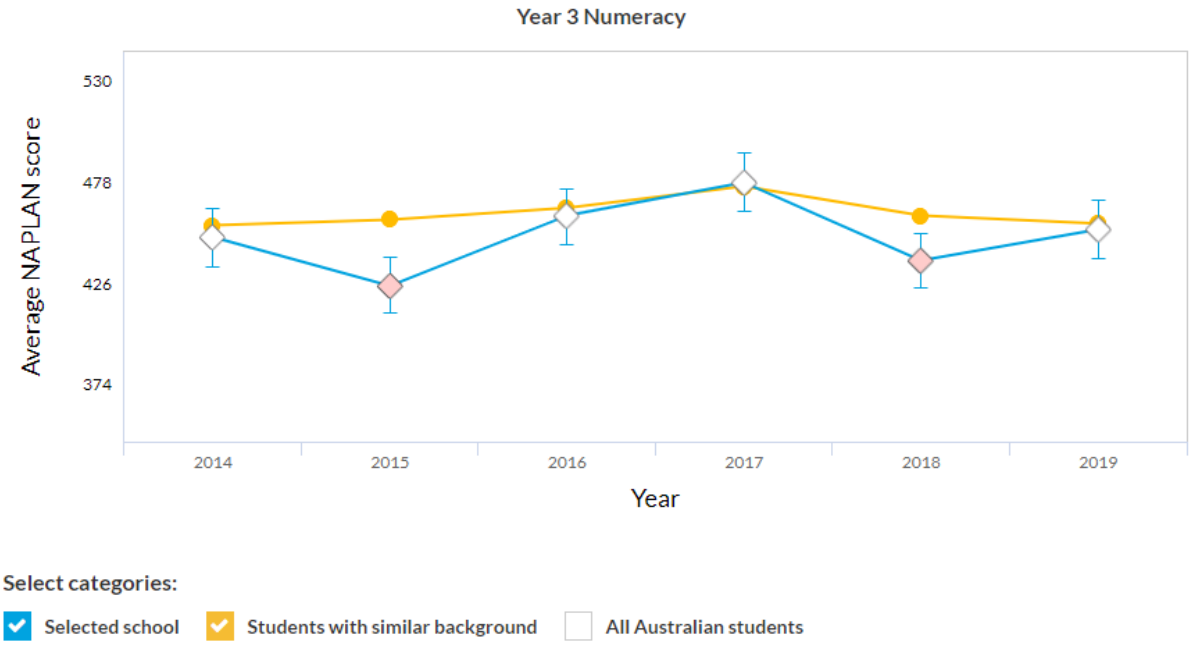
The adjusted SEA index was included in both newly developed analysis models for *My School*: the *similar-students analysis* and the *student progress rate analysis*. These new models are described in the following two sections, with technical details from these analyses included as appendices. The adjustments to the SEA model were used specifically for these analyses. The original 2013 model was applied when estimating SEA quarters and the Index of Community and Socio-Educational Advantage (ICSEA) for each school as published on *My School*. The reason for this was that there may be implications for the model for funding schools which would require agreement from states, territories and the Australian Government Department of Education, Skills and Employment (AGDESE) prior to change being made.

3. Similar-students analysis

In December 2019, Education Council endorsed a revised methodology for benchmarking the performance of schools on the *My School* website. Instead of comparing a school's NAPLAN result against the average result of 60 similar schools as was done in previous years, it was benchmarked against the average NAPLAN score of students with a similar background. Student background was defined by parental occupation and education (socio-educational advantage, or SEA), identification as being of Aboriginal and/or Torres Strait Islander origin (ATSI), and remoteness of the school. The average NAPLAN score of similar students was equal to the predicted achievement for each school given their average SEA, the percentage of Indigenous students and the remoteness of the school (utilising a multi-level regression model). If a school's result was significantly above their predicted score, the school's result was coloured green; if it was significantly below their predicted score the school's result was coloured red.

Generally, the predicted score for each school was close to the average result of 60 similar schools. However, the change in methodology caused some results to appear anomalous on the *My School* website. All results include a level of uncertainty, because they are estimates of unknown true values. In theory, the line graph that compares a school's official average achievement and the achievement of similar students (e.g. see Figure 1) could show inconsistencies with colouring presented in the *My School* tables. Such an inconsistency is due to differences in the level of uncertainty associated with the results generated using each method. For example, we do not know a student's true ability in numeracy, we estimate it by administering a test with certain items on a certain day. Had we used other items, or administered the test on another day, the student may have received a different achievement score. Both scores are estimates of the same student's ability. The uncertainty in *My School* results is larger for smaller schools, for schools with more variation in achievement between students, for schools with very low or very high achievement compared to the national average and for shorter tests (such as conventions of language and writing).

Figure 1: Example of a graph on My School showing results for similar student



For those schools with a very small number of students the level of uncertainty is particularly difficult to estimate. For this reason, schools with less than five students were excluded from the analysis and results for schools between five and ten students (inclusive) were coloured grey. In other cases, where anomalies appeared large, the consensus by NADAR was to colour similar student results in the NAPLAN table for 2019 white or, in extreme cases, to supress the average of the similar students. In total, 11 per cent of all similar-students results were coloured grey and 0.4 per cent (349 out of 92,263 results) were suppressed or coloured white due to uncertainty in the results.

Appendix B includes technical details of the statistical model and treatment of results.

4. Student progress rate analysis

Until NAPLAN 2018, *My School* reported school-level growth as the change in average achievement for students who took NAPLAN tests at the same school two years apart.

There has been considerable criticism over several years from stakeholders and technical experts that change in average achievement is of limited use when comparing growth between schools. Generally, schools with low levels of achievement have more space to grow than schools with high levels of achievement within the same period of time. Under the previous methodology for estimating growth, only those schools with low starting scores and high levels of growth are likely to be recognised for their change in achievement.

ACARA progressed the development of a new measure in consultation with external stakeholders. For this measure, a multiple regression analysis technique was used, with progress rates presented as the percentage of students within the school who achieved above the average growth of students who had both the same NAPLAN score two years ago and the same SEA score. In other words, the percentage of students in a school showing above average growth accounting for the school's average performance two years ago and the school's average SEA. Consequently, the chance of being a school with above average growth was independent of the level of achievement two years ago or the level of parental occupation and education.

Under this model, 50 per cent of Australian students will show above average growth and 50 per cent below. The percentage of students demonstrating above average growth within a school was compared with the national percentage of 50. As with every result on *My School*, each school percentage included a degree of uncertainty. As such, the difference between each school's percentage and the national percentage was tested for statistical significance. In cases in which a result was statistically significant and the school percentage was above 65 or below 35, the result was coloured dark green or dark red respectively. All other statistically significant results were coloured a lighter green if significantly above average or lighter red if significantly below average, i.e. either smaller than 46 per cent or larger than 54 per cent.

More technical details of the statistical model and procedures used to estimate student progress rates and of the treatment of results are included in Appendix C.

5. Appendices with technical details

Appendix A: Technical details of SEA model

Item response theory (IRT) was utilised to generate student scores on the socio-educational advantage (SEA) index. More specifically, the generalised partial credit model as implemented in ACER ConQuest V4 (Adams, Wu, Wilson, 2015) was used to estimate model parameters. The measurement model consists of two parts: the item model and the population model. In the item model, two parameters were estimated for each item: the item location (ξ_i) and the score parameter (τ_i). Unlike the more restricted one parameter partial credit model (Masters, 1982), the score parameter gives a different weight to each item. Full details of the model are documented in the online technical note by Macaskill & Adams (2016). For the population model, plausible value methodology was applied (Mislevy & Sheehan, 1987).)

The item model was estimated in the first step. The items are the manifestations of the construct being measured and therefore define the construct SEA. The SEA index was measured by eight items, four for each parent:

1. School education (SE)
    a. Year 9 or less
    b. Year 10
    c. Year 11
    d. Year 12
2. Non-school education (NSE)
    a. No non-school education
    b. Certificate I-IV, incl. trade certificate
    c. Advanced diploma / Diploma
    d. Bachelor's degree or above
3. Occupation group (OCC)
    a. Machine operator
    b. Trades person / Clerk / Sales
    c. Professional / Manager
    d. Senior Manager
4. Non-paid work (ONP)
    a. Paid work
    b. Non-paid work

Of the eight items used to measure SEA for each student, six were partial credit items with a maximum score of 3 and two were dichotomous items (1/0). Data from all jurisdiction and all NAPLAN year levels were included and senate weights were applied, ensuring each jurisdiction contributed equally to the item calibration process.

The EAP/PV reliability of the item model was 0.74. Location and score parameters are included in Table 1.

Table 1: Location and scoring parameters of the SEA model

Item	Location estimate (Xsi)	Scoring estimate (Tau)	Score 1	Score 2	Score 3
P1SE	-1.528	1.196	1.196	2.392	3.589
P2SE	-1.253	1.023	1.023	2.046	3.068
P1NSE	-0.258	1.226	1.226	2.453	3.679
P2NSE	-0.195	1.752	1.752	3.505	5.257
P1OCC	-0.020	0.998	0.998	1.996	2.994
P2OCC	0.029	1.247	1.247	2.494	3.741
P1ONP	-1.412	0.908	0.908	-	-
P2ONP	-2.556	1.086	1.086	-	-

The population model was estimated in the second step while anchoring the item parameters to the values estimated in the first step. Plausible value methodology was used to generate SEA values for each student. This methodology was chosen in 2013 because of its superior treatment of missing data compared to other methodologies investigated at that time. Instead of generating a point estimate for each student's SEA, this methodology estimates a range for each student that included their most likely SEA value (also called the posterior distribution; Wu, 2005; Monseur & Adams, 2009). Plausible values are random draws from this most likely range for each student. Five values were drawn for each student.

The *location* of the most likely range for each student's SEA is largely determined by the item responses (i.e. responses to parental occupation and education items). Having parents in high education and occupation groups resulted in SEA values at the top of the scale; having parents with low levels of education and low-income occupation groups resulted in SEA values at the bottom of the scale.

The *width* of the most likely SEA range is related to the uncertainty in the estimation. More uncertainty results in a wider range. Eight items are a small number of items and result in wider ranges than measurements based on more items (such as the NAPLAN test). In addition, the large number of missing values in parental occupation and education increased the uncertainty and therefore the width of the most likely SEA range for many students. Increasing the amount of information about the students can improve the precision of the estimates and therefore reduce the width of the most likely range from which plausible values are drawn. Information about students is added as regressors to the IRT model predicting the SEA construct. The relationship between the regressors and SEA is used to improve the estimation of the most likely ranges and especially improves the estimation of SEA for students without parental occupation and education data. It is important to note that adding information in the form of regressors does *not* change the meaning of the construct being measured.

Regressors used in the SEA model were NAPLAN reading achievement (weighted likelihood estimate), ATSI, missing ATSI information and geolocation (five categories dummy coded into four variables). In 2019, one additional regressor was added to take into account the large amount of missing data and the relatively low average SEA in the Northern Territory. Adding an indicator to identify students from the Northern Territory improved the estimation of SEA values for these students, especially for students without parental education and occupation data. This can be explained using a simplified model as an example. In a model with only item responses and no regressors, the best estimate for students without any parental data (missing responses to all eight items) is the national mean. This national mean is likely to be too high for the 12 per cent of students in the Northern Territory without parental data, because the average SEA in NT is more than half a standard deviation lower in the Northern Territory than in the rest of the country. By including a regressor for the Northern Territory the SEA plausible values for students from the Northern Territory without parental data will be randomly distributed around the jurisdictional mean for SEA instead of the national mean SEA, which is a better estimate for these students given the available information about them (i.e. they are from the Northern Territory).

Adding these regressors to the model improved the EAP/PV reliability from 0.74 to 0.79 (or 0.78 without the indicator for the Northern Territory). As a group, the regressors explained 26% of the variation in SEA. Table 2 shows that the width of the most likely ranges was smallest for the final model and largest for the model without any additional information about students (i.e. no regressors), suggesting that uncertainty in the estimates decreased by adding information about students to the model. It also shows that the uncertainty in the SEA estimates was largest for the Northern Territory. A likely explanation for this was the large amount of missing parent data for students in the Northern Territory (12% compared to 5% nationally).

Table 2: Average width in logits of most likely SEA ranges (or average standard deviation of posterior distributions) by jurisdiction for three SEA models

	Final 2019 model	2013 model (no NT)	No regressors
ACT	0.46	0.46	0.49
NSW	0.44	0.44	0.46
NT	0.49	0.50	0.52
QLD	0.44	0.45	0.47
SA	0.45	0.45	0.48
TAS	0.44	0.45	0.47
VIC	0.43	0.43	0.45
WA	0.45	0.46	0.49

Table 3 shows the resulting mean SEA and the standard deviation within jurisdictions and nationally. Mean SEA ranged from -0.64 in the Northern Territory to 0.48 in the ACT. Most other jurisdictions had a mean SEA near the national mean, except for TAS which had a mean SEA of -0.29. Variation in SEA was similar across states, except for the Northern Territory where the variation was larger.

Table 3: SEA mean and standard deviation by jurisdiction and nationally

	Mean SEA	SD of SEA
ACT	0.48	0.92
NSW	0.03	1.00
NT	-0.64	1.19
QLD	-0.05	0.93
SA	-0.03	0.93
TAS	-0.29	0.95
VIC	0.06	0.98
WA	0.01	0.95
Australia	0.01	0.98

Appendix B: Technical details of similar-students analysis

The methodology developed for *My School* 2019 was similar but somewhat different from the methodology used in previous years. In previous years, a school's average NAPLAN result was compared to the achievement of 60 schools with the most similar ICSEA scores. The current methodology can be regarded as comparing a school's average achievement with the average achievement of Australian students with a similar background as the students in that school. The average achievement of students with a similar background is equal to the predicted score from a regression model which is described in detail in this technical appendix.

The model

A multi-level regression model (MLM) with a school and a student level was applied to predict NAPLAN scores from the ICSEA components of socio-educational advantage (SEA), Indigenous status (ATSI) and remoteness of the school (ARIA). This model was similar to the multi-level model used to generate ICSEA scores for schools, as described in the ICSEA 2013: Technical report.

For a random intercept, fixed slopes MLM, the system of equations was

Level 1 (student)

$$Y_{ij} = \beta_{0j} + \beta_1 SEA_{ij} + \beta_2 ATSI_{ij} + \beta_3 ATSI_{ij}^{mis} + r_{ij}$$

Level 2 (school)

$$\beta_{0j} = \gamma_{00} + \gamma_{01} SEA_j + \gamma_{02} ATSI_j + \gamma_{03} ARIA_j + u_{0j}.$$

At level 1, SEA_{ij} was the SEA value for student *i* in school *j*, $ATSI_{ij}$ the Indigenous status of student *i* in school *j*, $ATSI_{ij}^{mis}$ an indicator for missing ATSI information for student *i* in school *j* and r_{ij} was the student-level residual. At level 2, SEA_j was the mean SEA score of students in the same year level of school *j*, $ATSI_j$ the percentage of Indigenous students in school *j*, $ARIA_j$ the remoteness of school *j* and u_{0j} the school-level residual for school *j*.

The system of two equations described above can be rewritten as a mixed effect model by substituting β_{0j} in the equation for level 1. This results in the following full model.

$$Y_{ij} = \gamma_{00} + \gamma_{01}SEA_j + \gamma_{02}ATSI_j + \gamma_{03}ARIA_j + \gamma_1 SEA_{ij} + \gamma_2 ATSI_{ij} + \gamma_3 ATSI_{ij}^{mis} + u_{0j} + r_{ij}$$

Of particular interest for the reporting of similar-student analysis on *My School* were the school-level residuals u_{0j} and their associated standard errors $\sigma_{u_{0j}}$. The residual was the difference between the predicted school mean given the background of the students and the remoteness of the school attended by the student and the actual school mean. Positive residuals indicated higher achievement than predicted; negative residual indicated lower achievement than predicted.

The *lme4* package (Linear Mixed-Effects Models using 'Eigen' and S4; Bates et al. (2014)) from R was used for fitting the mixed-effects models and estimating the school level residuals and their standard errors. Restricted maximum likelihood (REML) was used to estimate variance belonging to random effects while simultaneously removing fixed effects.

Both the student NAPLAN scores and the SEA values were sets of five plausible values for each student. Therefore, each model was run five times, once for each plausible value. Results from the five models were combined to derive the final estimates. The final residual was the average of the five residuals for each school.

$$u_{0j} = \frac{1}{5}\left(u_{0j1} + u_{0j2} + u_{0j3} + u_{0j4} + u_{0j5}\right)$$

Similarly, the final error variance was the average of the five error variances.

$$\sigma^2_{u_{0j}} = \frac{1}{5}\left(\sigma^2_{u_{0j1}} + \sigma^2_{u_{0j2}} + \sigma^2_{u_{0j3}} + \sigma^2_{u_{0j4}} + \sigma^2_{u_{0j5}}\right)$$

This error variance, however, did not include measurement variance which could be estimated using plausible values. Measurement variance was calculated as the sum of the squared deviations of the five residuals from the mean residual, adjusted for the number of plausible values.

$$\sigma^2_m = \frac{1}{4} * \sum_{k=1}^{5}\left(u_{0jk} - u_{0j}\right)^2$$

The final error variance for each estimate was computed as the combination of the two error variances, with adjustment for the number of plausible values.

$$\sigma^2_{(error)} = \sigma^2_{u_{0j}} + \left(1 + \frac{1}{5}\right)\sigma^2_m$$

The correct standard error for each domain-specific school residual was thus the square-root of the final error variance.

$$\sigma_{(error)} = \sqrt{\sigma^2_{(error)}}$$

Method

At the data processing stage, the following students and schools were excluded from the analysis:

- students who did not sit the test
- students with a raw score of 0
- home schooled students or school ID equal to 0
- students in special schools
- school results based on less than 5 students.

Determining whether the actual school mean was significantly different from the predicted school mean was the same as testing if the residual was significantly different from zero. Using a significant level $\alpha=0.10$, confidence intervals were built around the residuals. The confidence interval was equal to the residual plus and minus 1.64 times the standard error. Results were statistically significant if the confidence interval did not include zero. This process was equivalent to dividing the residual by its standard error and comparing these z-scores with the criteria 1.64 and -1.64. Significantly positive residuals indicated achievement above predicted given the school's scores on the ICSEA components; significantly negative residuals indicated achievement below predicted.

The *My School* website uses colours to indicate whether schools performed as expected or not and if the difference was large or small. Some schools were too small to reliably calculate standard errors of residuals hence their results were coloured grey. Colouring of the significant results was based on effect size. The effect size was determined by comparing the standardised residual (school residual minus national mean of residuals, divided by the national standard deviation of the residuals) with a value of plus or minus 1.64. This would result in approximately 5 per cent dark red and 5 per cent dark green results. The following criteria were used for colouring results of the similar-students analysis:

1) dark green if statistically significant and standardised residual larger than 1.64
2) light green if statistically significant and standardised residual between 0 and 1.64 (or if not statistically significant but standardised residual larger than 1.64)
3) white if statistically not significant
4) light red if statistically significant and standardised residual between 0 and -1.64 (or if not statistically significant but standardised residual smaller than -1.64)
5) dark red if statistically significant and standardised residual smaller than -1.64
6) grey if school result was based on 10 or less students but more than four students.

Generally, the predicted score for each school was close to the average result of 60 similar schools as used in previous years. However, the change in methodology caused some results to appear anomalous on the *My School* website. All results include a level of uncertainty, because they are estimates of unknown true values. In theory, the line graph that compares a school's official average achievement and the achievement of similar students (e.g. see Figure 1) could show inconsistencies with colouring presented in the *My School* tables. Such an inconsistency is due to differences in the level of uncertainty associated with the results generated using each method.

The residuals from the regression model were equal to the difference between the school's actual mean calculated from plausible values and the school's predicted mean. The official school mean that was published on *My School*, however, was calculated from student weighted likelihood estimates (WLE), which are—unlike plausible values—discrete point estimates of student achievement. In most cases, the school mean calculated from plausible values was very close to the official school mean calculated from WLEs. However, in cases where school means were associated with larger uncertainties, the difference between school means estimated using alternate methods could be noticeable or even substantial. The uncertainty in *My School* results was larger for smaller schools, for schools with more variation in achievement between students, for schools with very low or very high achievement compared to the national average and for shorter tests (such as conventions of language and writing). More investigations were planned for the near future to determine the advantages and limitations of each type of school mean.

Where results appeared anomalous due to the change in methodology, results were coloured white (not significant) or the predicted mean was supressed. Criteria for apparent anomalous results to be coloured white were:

1) result was red, but official school mean was higher than predicted mean
2) result was green, but official school mean was lower than predicted mean
3) result was dark green/dark red but difference between official school mean and predicted mean was less than 10
4) result was white but difference between official school mean and predicted mean was more than 35 (and difference between PV mean and WLE mean was more than 20)
5) result was light green/light red but difference between official school mean and predicted mean was less than 2 or more than 35 (and difference between PV mean and WLE mean was more than 20).

Criteria for apparent anomalous results to be supressed were:

6) predicted mean was higher than the school's mean NAPLAN score by more than 40 points but was lower than the PV mean (these were also identified under criterion 1 or were based on 10 or less students)
7) predicted mean was lower than the school's mean NAPLAN score by more than 40 points but was higher than the PV mean (these were also identified under criterion 2 or were based on 10 or less students)
8) difference between school mean and predicted mean was larger than the national standard deviation (and difference between PV mean and WLE mean was more than 40).

These criteria identified 349 out of 92,263 results (0.4%) as anomalous. Of these 349 results, 231 were coloured white and 118 were supressed. Of the 118 supressed results, 72 were based on 10 or less students. Furthermore, of the 349 identified results, 296 (85%) were results for the conventions of language assessments (spelling and grammar & punctuation).

Model assumptions

Linear regression models assume linearity of relationships and homoscedasticity of residuals. A correlation of zero between the residuals and predicted scores would suggest these assumptions have been met. Figure 2 shows that there was some degree of violation of the model assumptions. This graph includes school results across all year levels and all domains (each dot is one school results). The correlation between the school level residuals and the predicted means was 0.08 (excluding results based on 10 or less students). Inspection of the schools in the bottom left of the graph and the top right of the graph revealed that there was some heteroscedasticity, or perhaps non-linearity, at the lower end for schools with a high percentage of ATSI students and at the higher end for fully selective high schools.

Adding quadratic and cubic relationships with SEA to the model and an indicator for the percentage of low SEA students (percentage of students in bottom quarter plus half the percentage in the second lowest quester) did not improve the degree of heteroscedasticity or non-linearity. Heteroscedasticity mostly affects the estimation of standard errors, not the regression coefficients. However, non-linearity or outliers could affect the regression coefficients and therefore the interpretation of residuals. However, the impact of this was regarded to be small and consequently it was decided to accept this weak correlation between residuals and predicted school means for NAPLAN 2019, with a view to research this phenomenon in greater detail for future NAPLAN cycles.

Figure 2: Scatterplot between school-level residuals and predicted school means of the similar-students analysis

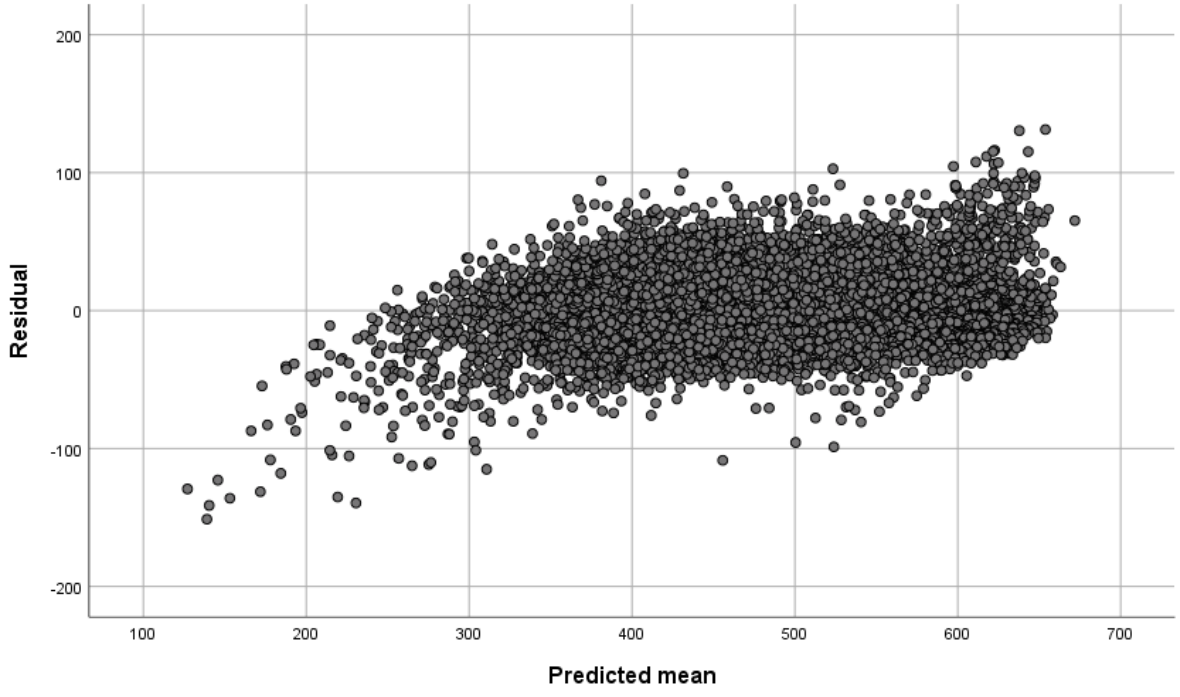
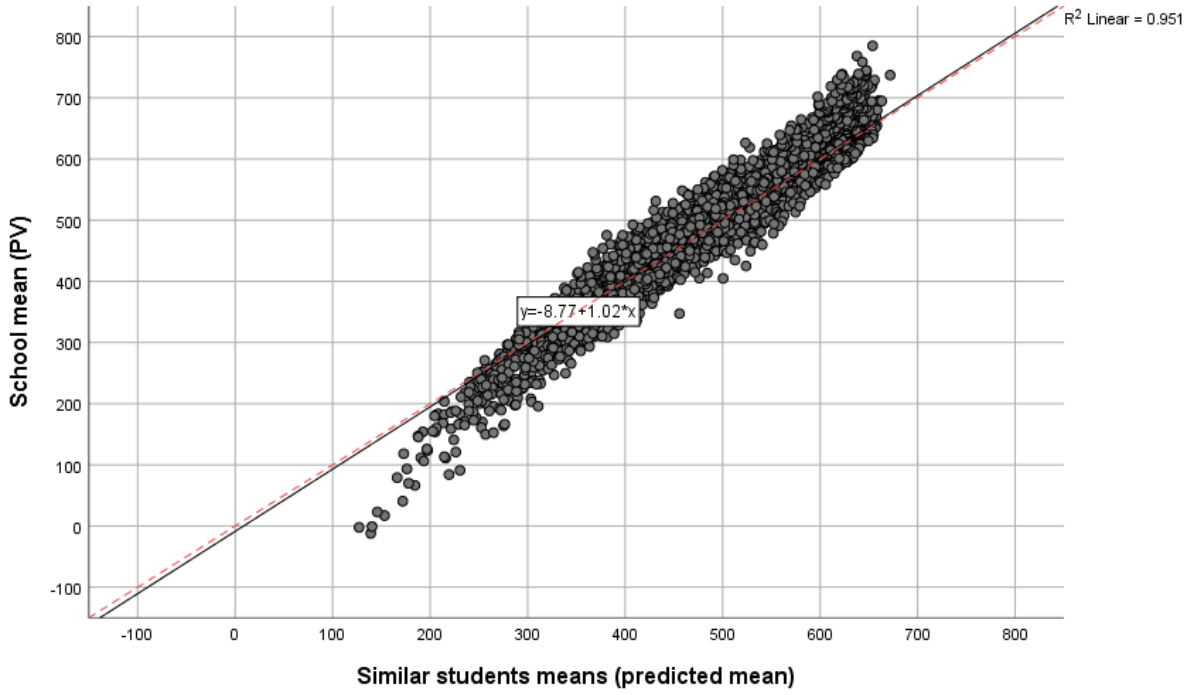


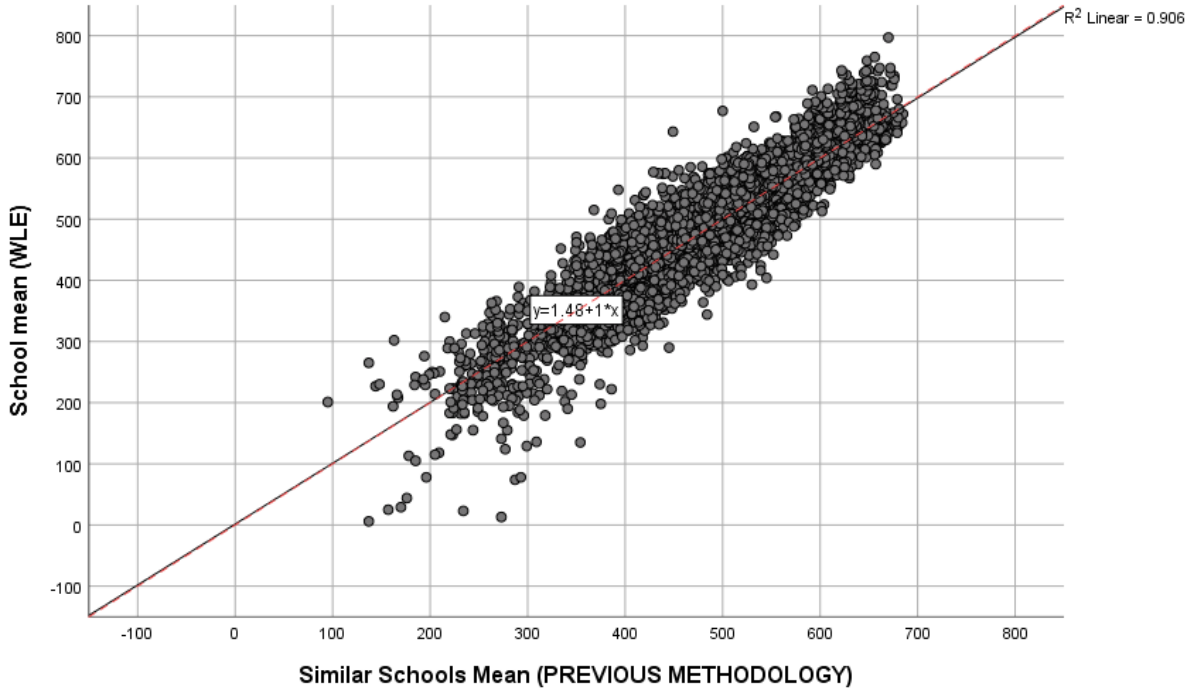
Figure 3 supports the finding described above. This scatterplot shows the relationship between the actual school mean (based on plausible values) and the predicted school mean (the average of similar students). All year levels and all domains are included in this graph. While the regression line (solid, black line) was very close to the identity line (broken, red line), schools with the lowest means (below 200 = 0.1% of all school results) performed below predicted and schools with the highest means (above 680 = 0.1% of all school results) performed above predicted. The predicted school means explained 95 per cent of the variation in actual school means.

Figure 3: Scatterplot between predicted mean (similar students) and the actual school mean (PV)



In contrast to the current approach, the average achievement of 60 similar schools from the previous methodology explained 91 per cent of the variation in actual school means (see Figure 4). The difference between the two R-squares suggested the new method was more accurate than the previous method. The degree of heteroscedasticity was similar between the two methodologies. That is, school means above approximately 680 were all above expected and school means below 200 were all below expected.

Figure 4: Scatterplot between the mean of similar schools and the actual school mean (WLE)



The correlation between the residuals of the similar-students approach and the residuals of the similar-schools approach (the difference between the school's achievement and the average achievement of the 60 similar schools) was 0.86 (excluding very small schools and anomalous results). The residuals of the new approach had a mean of 0 and a standard deviation of 15; the residuals of the previous approach had a mean of -1 and a standard deviation of 20. The difference in standard deviations supports the finding that the new approach is more accurate.

Results

Table 4 to Table 8 present regression coefficients of the MLM and their standard errors for each NAPLAN domain. All regression coefficients were statistically significant, except for ARIA. While the remoteness of the school is known to have a negative relationship with performance, this relationship disappeared once differences in SEA and Indigenous background were accounted for.

Table 4: Regression coefficients for the similar-students analysis of achievement in numeracy

		Year 3		Year 5		Year 7		Year 9	
	Coefficient	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Numeracy	***Intercept***	410	(0.3)	498	(0.3)	557	(0.5)	596	(0.5)
	Student SEA	24	(0.2)	22	(0.2)	23	(0.2)	18	(0.2)
	ATSI	-21	(0.6)	-20	(0.5)	-27	(0.6)	-20	(0.6)
	Missing ATSI	7	(1.3)	6	(1.2)	11	(1.3)	7	(1.1)
	School SEA	22	(0.6)	20	(0.5)	36	(0.8)	33	(0.9)
	% ATSI	-0.4	(0.02)	-0.3	(0.02)	-0.3	(0.03)	-0.1	(0.03)
	ARIA	-0.1	(0.17)	-0.4	(0.15)	1.1	(0.24)	0.9	(0.23)

Table 5: Regression coefficients for the similar-students analysis of achievement in reading

		Year 3		Year 5		Year 7		Year 9	
	Coefficient	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Reading	***Intercept***	434	(0.3)	510	(0.3)	550	(0.4)	587	(0.4)
	Student SEA	31	(0.2)	26	(0.2)	24	(0.1)	22	(0.2)
	ATSI	-19	(0.8)	-21	(0.6)	-21	(0.5)	-21	(0.7)
	Missing ATSI	5	(1.4)	8	(1.5)	7	(1.1)	4	(1.3)
	School SEA	24	(0.6)	20	(0.5)	29	(0.7)	30	(0.8)
	% ATSI	-0.4	(0.03)	-0.5	(0.02)	-0.3	(0.03)	-0.3	(0.03)
	ARIA	-0.4	(0.17)	-0.2	(0.14)	0.4	(0.20)	0.4	(0.21)

Table 6: Regression coefficients for the similar-students analysis of achievement in writing

		Year 3		Year 5		Year 7		Year 9	
	Coefficient	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Writing	***Intercept***	428	(0.3)	480	(0.3)	521	(0.5)	558	(0.6)
	Student SEA	15	(0.1)	16	(0.1)	17	(0.2)	19	(0.2)
	ATSI	-21	(0.5)	-20	(0.6)	-24	(0.6)	-26	(0.7)
	Missing ATSI	8	(1.0)	8	(1.3)	12	(1.2)	8	(1.4)
	School SEA	16	(0.5)	17	(0.6)	27	(0.8)	33	(1.1)
	% ATSI	-0.6	(0.02)	-0.5	(0.02)	-0.6	(0.03)	-0.6	(0.04)
	ARIA	-1.4	(0.15)	-1.6	(0.15)	-1.2	(0.22)	-0.1	(0.28)

Table 7: Regression coefficients for the similar-students analysis of achievement in spelling

		Year 3		Year 5		Year 7		Year 9	
	Coefficient	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Spelling	***Intercept***	421	(0.4)	505	(0.3)	550	(0.5)	588	(0.4)
	Student SEA	24	(0.2)	20	(0.2)	18	(0.2)	17	(0.2)
	ATSI	-20	(0.7)	-17	(0.6)	-18	(0.6)	-18	(0.6)
	Missing ATSI	6	(1.5)	8	(1.4)	7	(1.2)	6	(1.2)
	School SEA	19	(0.7)	16	(0.6)	24	(0.8)	25	(0.8)
	% ATSI	-0.4	(0.03)	-0.4	(0.02)	-0.3	(0.03)	-0.3	(0.03)
	ARIA	-2.6	(0.19)	-2.4	(0.16)	-1.5	(0.21)	-1.3	(0.23)

Table 8: Regression coefficients for the similar-students analysis of achievement in grammar and punctuation

		Year 3		Year 5		Year 7		Year 9	
	Coefficient	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Grammar & punctuation	***Intercept***	443	(0.4)	504	(0.4)	547	(0.5)	580	(0.5)
	Student SEA	32	(0.2)	28	(0.2)	25	(0.1)	22	(0.2)
	ATSI	-25	(0.7)	-24	(0.7)	-28	(0.6)	-20	(0.6)
	Missing ATSI	10	(1.5)	6	(1.6)	10	(1.3)	5	(1.2)
	School SEA	28	(0.7)	24	(0.6)	31	(0.8)	31	(0.9)
	% ATSI	-0.5	(0.03)	-0.5	(0.03)	-0.4	(0.03)	-0.4	(0.03)
	ARIA	-0.8	(0.19)	-1.0	(0.16)	-0.4	(0.22)	-0.4	(0.25)

The amount of variance that was explained by the ICSEA components between school, within school and in total is listed in Table 9 by domain and year level. Of the total variation in NAPLAN scores, about one quarter was explained by the ICSEA components. Of the variation in NAPLAN school means, the ICSEA components explain around 80 per cent. The ICSEA components explain 10 per cent of the variation in student achievement within schools.

Table 9: Percentage of variance in NAPLAN achievement explained by ICSEA components

Domain	Year level	Between schools	Within Schools	Total
Numeracy	Y3	78	10	26
	Y5	77	11	27
	Y7	82	10	32
	Y9	76	9	29
Reading	Y3	81	12	28
	Y5	83	13	31
	Y7	87	12	34
	Y9	84	11	33
Writing	Y3	71	6	24
	Y5	76	6	24
	Y7	82	6	30
	Y9	77	6	27
Spelling	Y3	71	7	20
	Y5	74	7	21
	Y7	79	6	24
	Y9	77	6	24
Grammar	Y3	80	11	29
	Y5	81	12	30
	Y7	86	12	36
	Y9	82	10	34

Distributions of the school level results as presented on *My School* are included in Table 10 to Table 14. The percentages in the first column under each year level were percentages of all results, including grey results and anomalous results. The percentages in the second column were only of the red, white and green results, hence they add up to 100 per cent.

Table 10: Distribution of similar students results for numeracy

Numeracy	Year 3		Year 5		Year 7		Year 9	
	All	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*
Dark red	4	5	4	4	2	2	2	2
Light red	8	9	10	11	16	17	20	22
White	64	73	62	71	58	64	54	59
Light green	7	8	8	9	11	12	13	14
Dark green	5	5	5	6	4	5	4	4
Total	87	100	87	100	91	100	91	100

Table 11: Distribution of similar-students results for reading

Reading	Year 3		Year 5		Year 7		Year 9	
	All	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*
Dark red	4	5	4	4	3	3	3	3
Light red	7	8	6	7	14	15	15	16
White	66	76	67	77	63	69	60	66
Light green	6	7	6	7	8	9	10	11
Dark green	4	5	4	5	4	4	4	4
Total	87	100	87	100	92	100	91	100

Table 12: Distribution of similar-students results for writing

Writing	Year 3		Year 5		Year 7		Year 9	
	All	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*
Dark red	4	5	4	4	3	4	4	4
Light red	11	13	9	11	17	18	19	21
White	56	65	62	71	52	57	46	51
Light green	12	14	8	10	15	16	19	21
Dark green	4	4	4	5	4	5	3	4
Total	87	100	87	100	91	100	91	100

Table 13: Distribution of similar-students results for spelling

Spelling	Year 3		Year 5		Year 7		Year 9	
	All	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*
Dark red	4	4	3	4	3	3	2	3
Light red	9	10	8	9	14	15	16	18
White	62	71	64	74	60	66	57	63
Light green	8	10	7	8	10	11	11	12
Dark green	5	5	5	6	4	5	5	5
Total	87	100	87	100	91	100	91	100

Table 14: Distribution of similar-students results for grammar and punctuation

Grammar	Year 3		Year 5		Year 7		Year 9	
	All	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*	*All*	*R/W/G*
Dark red	4	4	3	4	3	3	2	2
Light red	8	10	8	10	15	16	17	19
White	63	73	63	74	60	66	55	61
Light green	7	8	7	8	10	11	12	14
Dark green	4	5	4	5	4	5	4	4
Total	86	100	86	100	91	100	90	100

Appendix C: Technical details of student progress rate analysis

Until NAPLAN 2018, *My School* showed a school's student growth by reporting the change in average achievement for students who took NAPLAN tests at the same school two years apart.

There has been considerable criticism over several years from stakeholders and technical experts that change in average achievement is of limited use when comparing growth between schools. Generally, low achieving schools have more space to grow than high achieving schools within the same period of time. Under the previous methodology for estimating growth, only schools with low starting scores that achieve high growth off this low base, were likely to be recognised for their change in achievements.

ACARA progressed the development of a new measure in consultation with DSG, NADAR and the chair of MAG. A multiple regression analysis technique was used, and progress rates were presented as the percentage of students at the school who achieved above the average growth of students who had the same NAPLAN score two years ago and the same SEA score. In other words, the percentage of students in a school showing above average growth took into account the school's average performance two years ago and the school's average SEA. Consequently, the chance of being a school with above average growth was independent of the level of achievement two years ago or the level of parental occupation and education.

The model

The statistical model used to estimate student progress rate was a multiple, linear regression model. In this model, current NAPLAN achievement was regressed on prior NAPLAN achievement (i.e. the student's achievement two year ago) and on student SEA. This analysis was conducted on numeracy, reading and writing achievement only.

The regression model can be formally written as:

$$Y_{ij} = \beta_0 + \beta_1 PRIOR_{ij} + \beta_2 SEA_{ij} + r_{ij}$$

where Y_{ij} is current NAPLAN achievement of student *i* in school *j*, $PRIOR_{ij}$ is NAPLAN achievement two years ago of student *i* in school *j*, SEA_{ij} the parental education and occupation value for student *i* in school *j* and r_{ij} the residual for student *i* in school *j*. Students performing as predicted received a residual near 0, students performing above predicted received positive residuals, students performing below predicted received a negative residual.

For each student the standardised residual was saved. The standardised residuals had a mean of 0 and a standard deviation of 1 and could be regarded as z-scores. These student level z-scores were transformed into probabilities using the standard normal distribution. This probability could be interpreted as the probability of performing above the average achievement of students with similar prior achievement and similar SEA or, in other words, above predicted. Students with a probability near 1 were most likely performing above predicted, while students with probabilities near 0 were most likely performing below predicted. This probability can be expressed as:

$$p_{ij} = Pr\left(z_{r_{ij}}\right)$$

Averaging these student probabilities within schools resulted in an average probability for each school indicating the proportion of students within each school achieving above the average achievement of students with the same starting score and the same SEA as the students in that school. Using this approach, the proportion is not simply a count of students who achieved above predicted but rather a weighted count based on the degree of their over- or underperformance. In other words, a student who achieved far above predicted was given more weight than a student who achieved only just above predicted. For *My School* reporting, the school level proportions were presented as percentages:

$$P_j = \frac{100}{n_j}\sum_{i=1}^{n_j} p_{ij}$$

where n_j is the number of students included in the analysis for school *j*. For the progress rate analysis, discrete point estimates were used for current and prior student achievement (WLE) and five plausible values were used for the SEA of each student. Hence, the regression analysis was conducted five times, once for each SEA plausible value, and the final school percentage for reporting was calculated as the average of the five percentages:

$$P_j = \frac{1}{5}\left(P_{j1} + P_{j2} + P_{j3} + P_{j4} + P_{j5}\right)$$

As with every result on *My School*, the school's percentage included a degree of uncertainty. The error variance of the percentage P_j was estimated by

$$\sigma^2_{P_j} = \frac{100}{n_j}\sum_{i=1}^{n_j} p_{ij}\left(1 - p_{ij}\right)$$

Similar to the school percentage, the final error variance was the average of the five error variances calculated for each plausible value.

$$\sigma_{P_j}^2 = \frac{1}{5}\left(\sigma_{P_{j1}}^2 + \sigma_{P_{j2}}^2 + \sigma_{P_{j3}}^2 + \sigma_{P_{j4}}^2 + \sigma_{P_{j5}}^2\right)$$

This error variance, however, did not include measurement variance which could be estimated using plausible values. Measurement variance was calculated as the sum of the squared deviations of the five school percentages from the mean percentage, adjusted for the number of plausible values.

$$\sigma_m^2 = \frac{1}{4} * \sum_{k=1}^{5} \left(P_{jk} - P_j\right)^2$$

The final error variance for each percentage was computed as the combination of the two error variances, with adjustment for the number of plausible values.

$$\sigma_{(error)}^2 = \sigma_{P_j}^2 + \left(1 + \frac{1}{5}\right)\sigma_m^2$$

The correct standard error for each domain-specific school percentage was thus the square-root of the final error variance.

$$\sigma_{(error)} = \sqrt{\sigma_{(error)}^2}$$

Method

At the data processing stage, the following students and schools were excluded from the analysis:

- students who did not sit the test in either assessment year
- students with a raw score of 0 in either assessment year
- home schooled students or school ID is 0
- students in special schools
- school results based on less than 5 students
- students who changed schools (these students were included in the regression analysis but excluded when calculating school results).

By default, 50 per cent of Australian students showed above average growth and 50 per cent below. The school's percentage of students showing above average growth was compared with the national percentage of 50. Using a significant level α=0.10, confidence intervals were built around the percentage. The confidence interval was equal to the percentage plus and minus 1.64 times the standard error. Results were statistically significant if the confidence interval did not include 50. Percentages significantly above 50 indicated achievement above predicted given the school's prior performance and the SEA level; percentages significantly below 50 indicated achievement below predicted.

The *My School* website used colours to indicate whether schools performed as expected or not and if the difference was large or small. Some schools were too small to reliably calculate standard errors of percentages hence their results were coloured grey. Based on previous research and confirmed for the current analysis model by Centre for Education Statistics and Evaluation (NSW) the minimal sample size required for Poisson binominal distribution was estimated as 10. The following criteria were used for colouring results of the progress rate analysis:

1. grey if school result was based on less than 10 but more than four students
2. dark green if statistically significant and percentage was larger than 65
3. light green if statistically significant and percentage was larger than 50 (they were all larger than 54) and smaller than or equal to 65
4. white if statistically not significant
5. light red if statistically significant and percentage was smaller than 50 (they were all smaller than 46) and larger than 35
6. dark red if statistically significant and percentage was less than 35

Model assumptions

Linear regression models assume linearity of relationships and homoscedasticity of residuals. The correlation between the predicted school mean (the average of the predicted NAPLAN scores for students in a school) and the average school residual across all domains and year levels was close to zero (0.04) and the pattern in the residual plot appeared random (see Figure 5), suggesting a linear relationship between dependent and independent variables. However, the scatterplot in Figure 5 shows some signs of heteroscedasticity with smaller variation in the residuals for schools with higher predicted scores. In other words, prior achievement and SEA explained more variation at the higher end of the scale than at the lower end. This may have a small effect on the parameter estimates but would have an effect on the estimation of standard errors. For NAPLAN 2019 it was decided to accept these limited violations of the assumptions of the regression model, but more research on this phenomenon was planned for the near future.

Figure 5: Scatterplot between school average residuals and predicted school means by the progress rate analysis

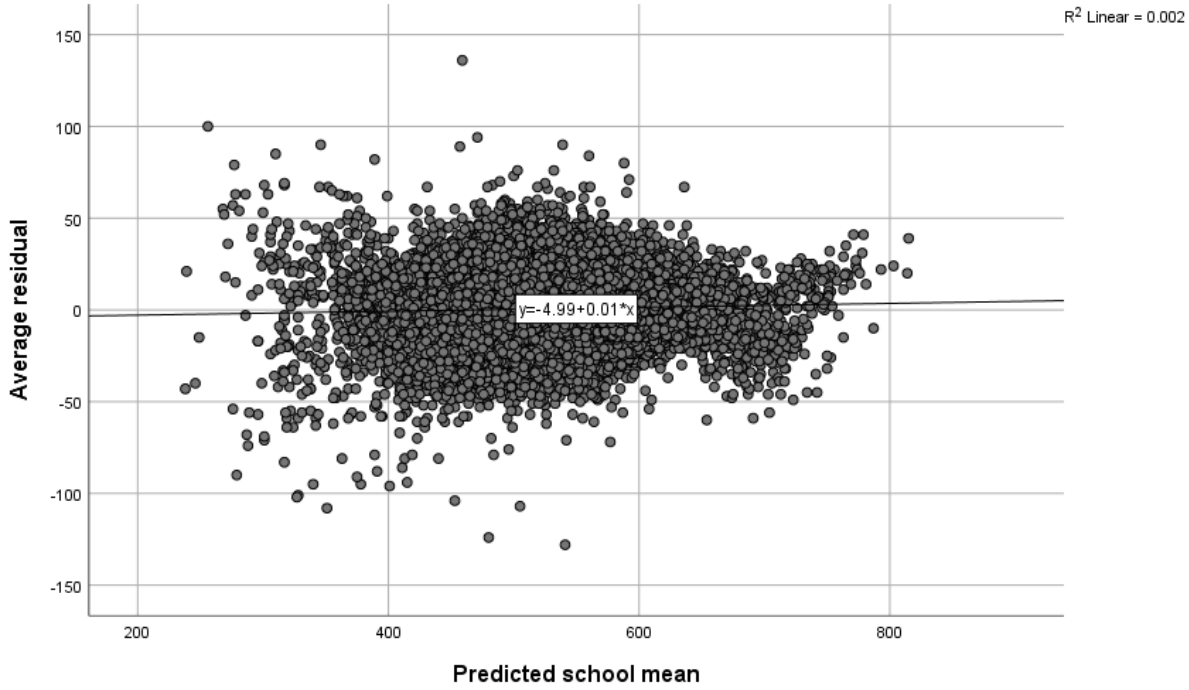
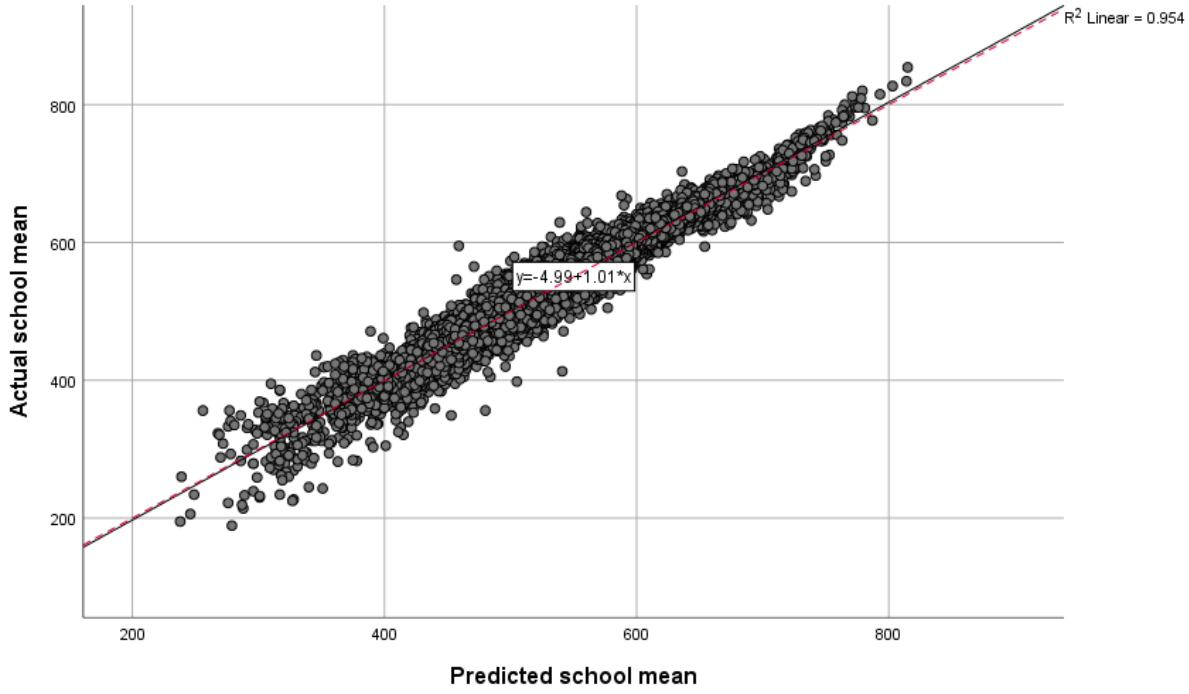


Figure 6 shows that the regression line is very close to the identity line and that the dots are evenly distributed around the identity line. As in Figure 5, the dots are somewhat wider spread at the lower end of the scale than the higher end of the scale.

Figure 6: Scatterplot between actual school means of matched students and the predicted school means by the progress rate analysis



Results

Regression coefficients and their standard errors are included in Table 15 by year level and domain. Both independent variables had a significant effect on current NAPLAN achievement. The two variables explained two-thirds (Year 3 to Year 5) to three quarters (Year 7 to Year 9) of variation in numeracy achievement (R-square), around 60 per cent of variation in reading and around 40 per cent of variation in writing.

Table 15: Regression coefficients and R-squares of the progress rate analysis by year level and domain

Progress	Coefficient	Numeracy		Reading		Writing	
		Estimate	*Std. Error*	*Estimate*	*Std. Error*	*Estimate*	*Std. Error*
Y3-Y5	Intercept	224	(0.5)	276	(0.5)	245	(0.7)
	Prior score	0.67	(0.001)	0.54	(0.001)	0.56	(0.002)
	SEA	9	(0.1)	14	(0.1)	11	(0.1)
	R-square	67%		59%		39%	
Y5-Y7	Intercept	114	(0.6)	237	(0.6)	216	(0.9)
	Prior score	0.89	(0.001)	0.61	(0.001)	0.63	(0.002)
	SEA	9	(0.1)	12	(0.1)	11	(0.1)
	R-square	73%		64%		40%	
Y7-Y9	Intercept	179	(0.6)	204	(0.7)	246	(1.0)
	Prior score	0.75	(0.001)	0.69	(0.001)	0.60	(0.002)
	SEA	5	(0.1)	9	(0.1)	12	(0.1)
	R-square	76%		63%		39%	

Distributions of the school level results of the progress rate analysis as presented on *My School* are included in Table 16 to Table 18. The percentages in the first column under each year level were percentages of all results, including grey results. The percentages in the second column were only of the red, white and green results, hence they add up to 100 per cent.

Table 16: Distribution of progress rate results for numeracy

Numeracy	Year 3 – Year 5		Year 5 – Year 7		Year 7 – Year 9	
Dark red	3	3	2	2	1	1
Light red	4	5	2	3	10	11
White	75	86	68	83	68	74
Light green	3	3	5	6	10	11
Dark green	2	3	5	6	3	3
Total	87	100	82	100	93	100

Table 17: Distribution of progress rate results for reading

Numeracy	Year 3 – Year 5		Year 5 – Year 7		Year 7 – Year 9	
Dark red	2	2	1	1	1	1
Light red	4	4	2	3	6	6
White	78	89	74	91	79	85
Light green	2	3	2	3	6	7
Dark green	2	2	2	3	1	1
Total	87	100	82	100	93	100

Table 18: Distribution of progress rate results for writing

Numeracy	Year 3 – Year 5		Year 5 – Year 7		Year 7 – Year 9	
Dark red	3	3	1	2	2	2
Light red	3	4	2	3	8	9
White	73	85	70	86	70	76
Light green	4	5	5	6	11	12
Dark green	3	3	3	4	2	2
Total	87	100	82	100	93	100

6. References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised Item Response Modelling Software* [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika 47*, 149–174.
Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of applied measurement, 10*(3), 1–15.

Mislevy, R. J., & Sheehan, K. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *The NEAP 1983-84 Technical Report, National Assessment of Educational Progress*. Princeton: Educational Testing Service.

Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of applied measurement, 10*(3), 320–334.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128.