

DEPARTMENT OF EDUCATION

Identifying potential strength and weakness in key learning areas using data from NAPLAN tests

Centre for Education Statistics and Evaluation



Centre for Education Statistics and Evaluation

The Centre for Education Statistics and Evaluation (CESE) was created in 2012 to improve the effectiveness, efficiency and accountability of education in New South Wales. It is focused on supporting decision-making in education delivery and development with strong evidence.

CESE analyses and evaluates educational programs and strategies and gauges New South Wales' education performance over time through its ongoing core data collections and delivery of analysis and reports. It also monitors national and international strategic agendas to ensure that New South Wales is well positioned to provide leadership in education.

CESE's three main responsibilities are:

1. to provide data analysis, information and evaluation that improve effectiveness, efficiency and accountability
2. to create a one-stop shop for information needs - a single access point to education data that has appropriate safeguards to protect data confidentiality and integrity
3. to build capacity across the whole education sector by developing intelligent tools to make complex data easy to use and understand, and providing accessible reports so that everyone can make better use of data.

CESE provides sound evidence for educators to make decisions about best practice in particular contexts and importantly, enables teachers to meet the needs of students at every stage of their learning.

Authors

Dr Lucy Lu, Dr Wai-Yin Wan.

Centre for Education Statistics and Evaluation, April 2019, Sydney, NSW

For more information about this report, please contact:

Centre for Education Statistics and Evaluation
Department of Education
GPO Box 33
SYDNEY NSW 2001

Email: info@cese.nsw.gov.au

Telephone: +61 2 7814 1527

Web: www.cese.nsw.gov.au

Acknowledgements

CESE would like to thank Professor David Andrich and Dr Ida Marais, University of Western Australia, for their valuable comments during the development of this report. The authors also thank the psychometric analysts in the Queensland Department of Education (James Cousin) and the Victorian Curriculum and Assessment Authority (Michael Dalton and Nathan Zoanetti) for productive discussions and their helpful advice.

CESE is entirely responsible for the content of the report.

Table of contents

1. Introduction	4
1.1 Report outline	5
2. Explanation of methodology and modelling process	6
2.1 Modelling process	6
2.2 Technical details of the measurement and statistical processes used	8
3. Quality of NAPLAN writing data – a prerequisite for skillset analysis	10
3.1 Item fit statistics and uni-dimensionality	10
3.2 Local independence	11
3.3. Structural validity	11
4. Minimum group size required when applying the proposed methodology	13
5. Model results using NAPLAN writing data (2013 – 2017)	15
5.1 Proportion of schools identified as weaker or stronger on a particular writing trait in a given year	15
5.2 Volatility of school-level model results	16
6. Validation of model results	18
6.1 External validity	18
7. Reporting options	20
7.1 Multiple-school report	20
7.2 Individual school trends report	20
7.3 Student level reports	21
7.4 Trial outcomes	23
8. Conclusions and next steps	24
References	25
Appendix A	
Marking rubric of NAPLAN writing test	26
Appendix B	
Item fit statistics – Infit and Outfit	27

1. Introduction

Federal, state and territory education ministers have agreed that NAPLAN will move online from 2018, with online tests also adopting an adaptive testing format rather than the traditional static testing format. Adaptive testing means that the test automatically adapts to a student's performance and presents questions that are appropriate for the student's achievement level. Adaptive testing provides more reliable and accurate information about high and low ability students as items can be better targeted to challenge and engage students throughout tests, soliciting performance that more accurately reflect students' underlying abilities. Once all schools have transitioned to online testing, it is also expected that the results will be delivered to schools quicker than the current paper tests, which means teachers potentially have more relevant test data to tailor their teaching specifically to student needs (ACARA, 2016).

The change of testing mode and format increases the complexity involved in teachers and school leaders appropriately using and interpreting NAPLAN data. There are three main challenges:

1. When using results from traditional paper-based tests, NSW teachers and school leaders often gauge the strengths and weaknesses in student performance by comparing the proportion of correct answers for a test item for a school or a class to the average in the state/ system, or to that in similar schools. When using results from the adaptive online tests, not all students are exposed to an item, meaning that the proportion of correct answers for an individual item will be less straightforward to interpret.
2. Online testing will result in each item being exposed to fewer students in a class; thus, focusing on students' performance on an individual item is likely to produce less reliable information than previous static tests.
3. Item content for the majority of online test items will not be released to teachers and school leaders.¹ Lack of visibility to the item content prevents deep analysis of item performance (e.g. distractor analysis²) that is traditionally undertaken by teachers in NSW when they receive NAPLAN data.

This paper details a new method of using NAPLAN test item data to inform teaching and learning. While new for NAPLAN, this method is similar to that used for analysing student performance patterns in Programme for International Student Assessment (PISA) (Yildirim, Yildirim & Verhelst, 2014).

The method represents a shift in the focus of test analysis from an individual item to a learning area or a skillset that is commonly assessed by a group of items. For each test (e.g. Year 3 reading, Year 5 numeracy), the process entails:

- first grouping all test items by the skillsets (or learning areas) the items assess, and then
- examining how students perform on a group of items assessing one common skillset, relative to students' overall performance in the domain.

Once we have identified particular skillsets where the performance of a group of students on the skillsets is better or worse than expected from robust measurement models, this information can then be provided to schools to help them identify teaching program strengths or weaknesses. By shifting the analysis focus from individual students on individual items to performance from a group of students on a set of items, the insight gained from the analysis will be more reliable and accurate. This analysis is referred to below as 'skillset analysis'.

¹ This is due to a heightened level of test security because online adaptive testing requires a much larger item bank than paper tests and therefore questions are more likely to be used across years. However, it is noted that the link between each item and the Australian Curriculum content areas will be provided to education systems, making the grouping of items by skillsets, and thus the application of the proposed methodology documented in this report, possible.

² Distractor analysis is the analysis of response patterns for incorrect options, for each item. It includes the investigation of possible reasons why students chose an incorrect answer.

Method summary

We use a generalised differential item functioning analysis approach to identify patterns of interest.

The analysis involves conducting Rasch modelling first to obtain person and item parameters, and then using a statistical process to evaluate whether a collection of responses (i.e. item scores) achieved by a group of students on a set of items assessing a common skillset is as expected or not based on this group of students' overall ability estimates for the test domain.

Rasch modelling (Rasch, 1960, 1980) is selected as it is underpinned by robust measurement principles and is commonly used for evaluating test validity. It is also used by the Australian Curriculum, Assessment and Reporting Authority (ACARA) for item evaluation and scale score computation for NAPLAN tests.

1.1 Report outline

This report explains the methodology, and the findings after applying this methodology, using writing test data as an example. Writing is an ideal domain to start with since NAPLAN writing is assessed using an analytic rubric across ten writing traits (or skillsets - see appendix A for the marking rubric used for NAPLAN writing). Each writing trait can readily be perceived as a conceptually distinct aspect of writing skill, and together the ten traits define the essential skills and knowledge required to produce an effective piece of writing.

The following sections illustrate how we may identify the areas or skillsets where students perform unexpectedly better or worse based on their overall performance in the whole domain.

Whilst this report uses writing as an example to illustrate the application of the proposed method, the method can be adapted to analyse test data from other domains such as reading and numeracy. Where appropriate, the report makes references to the adaption required in the illustrated methodology for other domains.

2. Explanation of methodology and modelling process

2.1 Modelling process

The methodology developed utilises the results from the NAPLAN tests. The NAPLAN tests are administered nationally to all students in Years 3, 5, 7 and 9. Students are assessed across five domains: reading, writing, spelling, grammar and punctuation, and numeracy.³ The aim of our method is to flag skillsets where student performance on items assessing a particular skillset are different from the model expectations.

The modelling process has a number of steps.

Identify the skillset

The process begins with the identification of key skillsets in a learning domain and grouping test items by them. For writing, there is no need to group items as each criterion forms a conceptually distinct skillset. For reading and numeracy, skillset identification has begun by the department's curriculum experts using a range of documents including National Literacy and Numeracy Learning Progressions (ACARA, 2018a, 2018b) and NSW English and Mathematics syllabuses (Board of Studies NSW, 2012a, 2012b).

Select the measurement model

The second step in the process is to identify a suitable Rasch model to calibrate the difficulty of the test item, and the ability of the student completing the test on a single measurement scale. This needs to be completed for each year cohort and for each of the reading, numeracy and writing domains.

For writing, the scripts are rated (polytomously)⁴ by markers using the same analytic rubric consisting ten traits (trait 1 to 10) across all four scholastic year levels. This means that a Polytomous Rasch model (Andrich, 1978) can be used, treating each trait as an item being scored using a rating scale with ordered categories. For reading and numeracy domains, a Rasch model (Rasch, 1960, 1980) is used as these tests contain only right/wrong (dichotomous) items (see Sections 2.2.1 and 2.2.2 for model technical specifications).

When developing measurement models, we use the whole of NSW government schools data. That means we compare the response patterns from a group of students on a particular set of items or a writing trait to the overall response patterns generated by all NSW government school students to all items/traits in that learning domain, taking into account the estimated overall ability of the target group of students.

Check that the model works well using diagnostic testing

As part of the psychometric modelling, necessary diagnostic checking such as local independence and uni-dimensionality and goodness of fit statistics are also performed to check whether the measurement model assumptions are satisfied. For each group of students of interest (e.g. a class or a school), model parameters and students' ability estimates are then used to calculate the probability of each student in that group receiving each of the possible outcomes for each of the items assessing a particular skillset. The details of the psychometric analyses will be given in the next section.

For polytomous items, an individual student's response on a particular item follows a categorical distribution with probability of the student scoring in each response category (referred to below as 'response probabilities') estimated from the psychometric analysis. Note that response probabilities vary across students based on their ability estimates (i.e., this categorical distribution is non-identical across students). The sum of scores received by a group of students on a set of polytomous items assessing a skillset can be considered as a sum of many categorical distributed random variables. For dichotomous items, the categorical distribution for an individual student's response on a particular item reduces to a Bernoulli distribution and the sum of multiple Bernoulli distributed random variables with differential response probabilities becomes a Poisson-binomial distribution. These summation processes of the

³ For more information about NAPLAN, see ACARA website: <http://www.nap.edu.au/naplan>.

⁴ Each trait is scored using a rating scale that includes successive integer score points (e.g. a 0-3 rating scale), with higher integers intended to indicate increasing levels of attainment for that trait. These traits are also referred to as polytomous items.

discrete distributed random variables can be computed using the R software package. Discussion on the minimum sample size required for the resulting distribution of the sum of scores approaching a normal distribution is in Section 4. The categorical random variables (or Bernoulli random variables) are non-identically distributed because each student has a different set of response probabilities (see Section 2.2.4).

Identify if a student's performance is statistically different from their expected pattern of achievement

With the asymptotic distribution of the sum of item scores (referred to as 'sum score') approaching normal, a hypothesis testing can be performed to test whether the observed sum score is higher or lower than what is expected from its empirical distribution. The null hypothesis is that there is no difference between the observed sum score and the expected sum score whereas the alternative hypothesis is that the observed sum score is greater than or smaller than the expected sum score. The test statistic is the observed sum score for all the items assessing a skillset from a group of students. The p-value, hereby referred to as the significance of occurrence (S_o), is the probability of observing the test statistic or more extreme given the known response probabilities. Depending on the magnitude of the significance of occurrence, students' performance on a skillset can then be classified into different groups, namely significantly above expectation, possibly above expectation, as expected, possibly below expectation and significantly below expectation.

Identify the extent of difference that is large enough to have practical implications

In addition to the statistical significance testing mentioned above, our proposed method also considers the practical significance of the difference between the observed sum score and the expected sum score for a collection of student-item scores. The idea is to flag only those differences that are worth following up by teachers as statistical significance testing is sensitive to sample size. One way of setting the practical significance level is to base it on the difference between the observed and expected sum scores, averaged across all items assessing a skillset for a group of students.

For NAPLAN writing analysis, a provisional threshold of average difference of 0.1 is used in the following analysis. For a given trait, this threshold is equivalent to 10% of all the trait scores, received by a group of students included in analysis, differing by 1 raw score point from their respective expected trait scores. Based on the standard deviations of trait scores from the NSW government students' population, this difference translates to an effect size of 0.1 to 0.2 across the ten traits. The practical threshold is preliminary and is subject to further fine tuning as we receive feedback from teachers and school leaders.

Categorise group performance based on thresholds for both statistical and practical significance

Based on the statistical significance (significance of occurrence) and the practical significance levels, the performance of a target group of students on a set of items assessing a skillset is classified into four groups: significantly above expectation, possibly above expectation, typical performance (also referred to as 'as expected'), possibly below expectation and significantly below expectation. Table 1 provides the decision rules used to classify the students' performance.

Table 1:

Classification of performance category using statistical significance and practical significance

Performance category	Colour label	Sum of scores in the skillset	Significance (S_o) limits	Average difference
Significantly above expectation		observed > expected	$S_o < 0.05$	≥ 0.2
Possibly above expectation		observed > expected	(i) $0.05 \leq S_o < 0.10$	≥ 0.1
			(ii) $S_o < 0.05$	$0.1 \leq d < 0.2$
Typical performance			Schools not classified in any of the 'above' or 'below' categories	
Possibly below expectation		observed < expected	(i) $0.05 \leq S_o < 0.10$	≥ 0.1
			(ii) $S_o < 0.05$	$0.1 \leq d < 0.2$
Significantly below expectation		observed < expected	$S_o < 0.05$	≥ 0.2

2.2 Technical details of the measurement and statistical processes used

2.2.1 Rasch model

Rasch model is a logit-linear model with the following specification:

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) = \beta_n - \delta_i \quad (1)$$

where P_{ni1} is the probability of student n to score 1 (answer correctly) on dichotomous item i . β_n is the latent ability measure (latent trait) of student n , and δ_i is the difficulty measure (or location) of item i on the same latent continuum which gives an indication of the β needed to correctly respond to the item. Equation (1) can be re-written such that the probability of student n to score 1 for item i is expressed as:

$$P_{ni1} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \text{ and } P_{ni0} = 1 - P_{ni1} \quad (2).$$

2.2.2 Polytomous Rasch model

In the case of NAPLAN writing test, all of the ten writing criteria are scored using more than two ordered categories resulting in a polytomous response dataset. Andrich (1978) defined the Polytomous Rasch model to analyse such data and expressed the model as:

$$P_{nix} = \frac{1}{\gamma_{ni}} \exp[-\sum_{k=0}^x \tau_{ki} + x(\beta_n - \delta_i)]; \quad x = 0, 1, \dots, m_i \quad (3)$$

where P_{nix} is the probability of student n to score x ($x=0, 1, \dots, m_i$) on polytomous item i ; m_i is the maximum score of item i ; β_n and δ_i are as defined in the previous section; τ_{ki} is the threshold parameter for item i which denotes the point at which the probabilities of responses in category k and category $k-1$ are equal. The threshold parameters are subject to two constraints:

$$\sum_{k=1}^{m_i} \tau_{ki} = 0 \text{ and } \tau_{0i} = 0.$$

γ_{ni} is a normalizing factor that ensures the probabilities in Equation (3) sum to 1:

$$\gamma_{ni} = 1 + \sum_{x=1}^{m_i} \exp[-\sum_{k=1}^x \tau_{ki} + x(\beta_n - \delta_i)].$$

Note that the Polytomous Rasch model is also known as the Partial Credit model and Dichotomous Rasch model is a special case of it. If all the items have the same thresholds, it is reduced to a Rating Scale model (Andrich, 2005). 'TAM' package in 'R (version 3.5.0)' was used to fit the Polytomous Rasch model in which Marginal maximum likelihood estimation (MMLE) method is used to estimate the item and person parameters.

2.2.3 Calculating expected value and variance for the trait score

Once the parameters of a psychometric model are estimated, they are then used to compute the expected response or score of every item for each student and the variance for the observed response. The expected response or score E_{ni} and the corresponding variance V_{ni} for student n and item i are calculated based on the response probabilities derived above using the following formula:

$$E_{ni} = \sum_{x=0}^{m_i} j \times P_{nix} \text{ and } V_{ni} = \left(\sum_{x=0}^{m_i} x^2 \times P_{nix}\right) - E_{ni}^2.$$

2.2.4 Summing individual categorical/Bernoulli distributions

This section illustrates the calculation of the probability mass function of the sum of the two categorical random variables for two students' scores for a writing trait. The sum of more than two categorical random variables can be computed using the same algorithm.

Let X be the discrete categorical random variable for the score of student A and Y be the discrete categorical random variable for the score of student B ; then the sum of the two writing scores can be defined as $Z = X + Y$ which is also a discrete random variable. Let the probability mass functions of the discrete categorical random variable X and Y be $P(X = x)$ and $P(Y = y)$ respectively for student A and B , then the joint probability mass function for the sum of the two writing scores can be specified as:

$$P(Z = z) = \sum_x P(X = x, Y = z - x)$$

The above calculation can be iterated by adding another discrete categorical random variable to the resulting sum of discrete categorical random variables. As the sample size increases, the resulting distribution approaches a normal distribution.

For dichotomous items, the same summing process can be applied to the Bernoulli distributed random variables. The resulting distribution of the sum of Bernoulli distributed random variables becomes a Poisson-binomial distribution. The probability of having k correct responses out of a total of n can be written as:

$$P(K = k) = \sum_{A \in F_k} \left(\prod_{i \in A} p_i \right) \left(\prod_{j \in A^c} (1 - p_j) \right)$$

where F_k is a set of all subsets of k integers that can be selected from $\{1, 2, 3, \dots, n\}$. For example, if $n = 3$, then $F_2 = \{\{1,2\}, \{1,3\}, \{2,3\}\}$. A^c is the complement of A .

R software (the TAM package) was used to run the measurement models and obtain the person and item parameters and the student-item response probabilities. It is also used to compute the resulting distribution from summing the individual random variable distributions. Another package 'discreteRV' was used to evaluate the sum of categorical distributed random variables and the package 'poibin' was used to compute the Poisson binomial distribution.

The following section provides findings from the checking of model assumptions underlying the Rasch modelling before presenting results from using this methodology when applied to writing.

3. Quality of NAPLAN writing data – a prerequisite for skillset analysis

An important step in the skillset analysis is to check that the quality of the test data is sufficient for the proposed statistical analysis. In order to achieve this, we investigated construct validity for the test data, including examining evidence of measurement validity, a type of check particularly important given the use of Rasch models in the analyses to establish expectations for comparisons to the observed data. Two key assumptions underpinning the Rasch measurement models outlined above are uni-dimensionality and item local independence.

3.1 Item fit statistics and uni-dimensionality

The first assumption is that all items from a test are assessing a single latent ability.

If this assumption is violated, it means that we can no longer be certain about the validity of the total test score as there is 'noise' in the measurement (e.g., due to construct irrelevant variance introduced into the measurement process). A key task of ACARA⁵ is to check and ensure that there is sufficient evidence that NAPLAN items are assessing a single construct, and that where response categories are intended to be ordered, there is evidence to support that assumption.

Given our analysis only uses NSW government school students' data, these assumptions are checked during our psychometric analysis through examination of item fit statistics, order of response category thresholds, and proportion of variance in the item scores explained by the Rasch dimension and secondary dimensions. Appendix B provides technical details of the types of item fit statistics used in our analysis.

Item fit and uni-dimensionality in writing results

Using writing data from 2013 to 2017, Rasch analysis shows that across the ten writing traits, all the infit and outfit statistics are within the reasonable range of 0.6 to 1.4 (Wright & Linacre, 1994). As an illustration, fit statistics using 2017 Year 3 test data are provided in Table 2. The table indicates that the ten writing traits fit a uni-dimensional model reasonably well. The only comment worthy of making is there is some evidence of Audience assessing writing features that are somewhat overlapping with those assessed by other traits.

Table 2:

Item-fit statistics from the Polytomous Rasch model for NAPLAN 2017 Year 3 writing

Marking criteria	Infit	Outfit
Audience	0.72	0.69
Text structure	0.94	0.94
Ideas	0.87	0.87
Persuasive devices	0.98	0.96
Vocabulary	0.85	0.85
Cohesion	0.85	0.81
Paragraphing	1.05	1.06
Sentence structure	0.93	0.93
Punctuation	1.10	1.11
Spelling	1.04	1.05

Examination of category thresholds for each trait shows all categories are properly ordered as intended for all traits.

⁵ ACARA is the national assessment and reporting agency responsible for the implementation of NAPLAN.

Further analysis shows that the Rasch dimension (item and person measures) explained 66% of the total variance in the data, higher than the proposed threshold of 50% (Linacre, 2009). The remaining 34% of total variance is unexplained and a Principal Component Analysis (PCA) was conducted on the standardised residuals. Results from the PCA show that the unexplained variance (Eigenvalue) in the first principal component is 1.83 which accounts for 6.2% of the total variance. Unexplained variance (Eigenvalue) in the second principal component is 1.40 which accounts for 4.8% of the total variance. These findings suggest no evidence of the violation of uni-dimensionality in the NAPLAN 2017 Year 3 writing test.

3.2 Local independence

The second assumption underpinning Rasch models is local independence, which assumes that a correct or wrong response to one item should not lead to a correct or wrong answer to another item (Hambleton & Swaminathan, 1985).

To check the local independence assumption, a common method is to examine the correlations between the items that are not accounted for by the latent trait (i.e. the person parameters). When the absolute value of residual correlation between a pair of items is greater than 0.3, we take it as an indication that there is a possible dependence between the pair of items (Christensen & Kreiner, 2013).

Table 3 shows the correlation matrix of the standardised residuals between pairs of writing traits. All pairs have a correlation between -0.3 and 0.3, indicating no significant evidence of the local independence assumption being violated. The only pair of correlation that in size is close to 0.3 is the negative correlation between Punctuation and Ideas and that between Spelling and Text Structure. These correlations reflect an interesting contrast existent in the writing construct domain, which is explored in the next section.

Table 3:

Correlation of the standardised residuals between pairs of writing traits

Marking criteria	Audience	Text structure	Ideas	Persuasive devices	Vocabulary	Cohesion	Paragraphing	Sentence structure	Punctuation	Spelling
Audience	1.000	-0.065	0.149	-0.036	-0.047	-0.061	-0.178	-0.135	-0.235	-0.142
Text structure		1.000	-0.091	0.168	-0.138	-0.151	0.045	-0.257	-0.225	-0.298
Ideas			1.000	0.011	-0.025	-0.078	-0.177	-0.168	-0.257	-0.193
Persuasive devices				1.000	-0.088	-0.134	-0.168	-0.203	-0.242	-0.275
Vocabulary					1.000	0.009	-0.097	-0.068	-0.122	-0.030
Cohesion						1.000	-0.084	-0.009	-0.116	-0.058
Paragraphing							1.000	-0.168	-0.116	-0.192
Sentence structure								1.000	0.004	-0.041
Punctuation									1.000	-0.007
Spelling										1.000

3.3. Structural validity

The internal structure of the trait scores is next analysed in order to further examine the quality of writing scores which may be susceptible to inconsistencies in marking. The focus here is to check whether trait scores exhibit internal patterns that are consistent with what is known about the trait being assessed, in particular the structural relations inherent in behavioural manifestations of the underlying construct (Messick, 1996). Loevinger (1957) refers to this type of evidence as evidence for 'structural validity'.

To do this, PCA was conducted on the trait residual scores (after the influence of the Rasch dimension is removed), using NAPLAN writing test data, for each of the four scholastic year levels and for each calendar year separately. Results are similar across calendar years, so only those from 2017 writing data are provided below. The factor loadings of traits on the first component are plotted against the year level in Figure 1.

It's clear from the graph that, consistently across the four scholastic years, the meaning of the first dimension (after the Rasch dimension) can be interpreted as the contrast between:

- word or sentence level writing features (sentence structure, punctuation and spelling), and
- whole text level writing features (audience, text structure, ideas and persuasive devices).

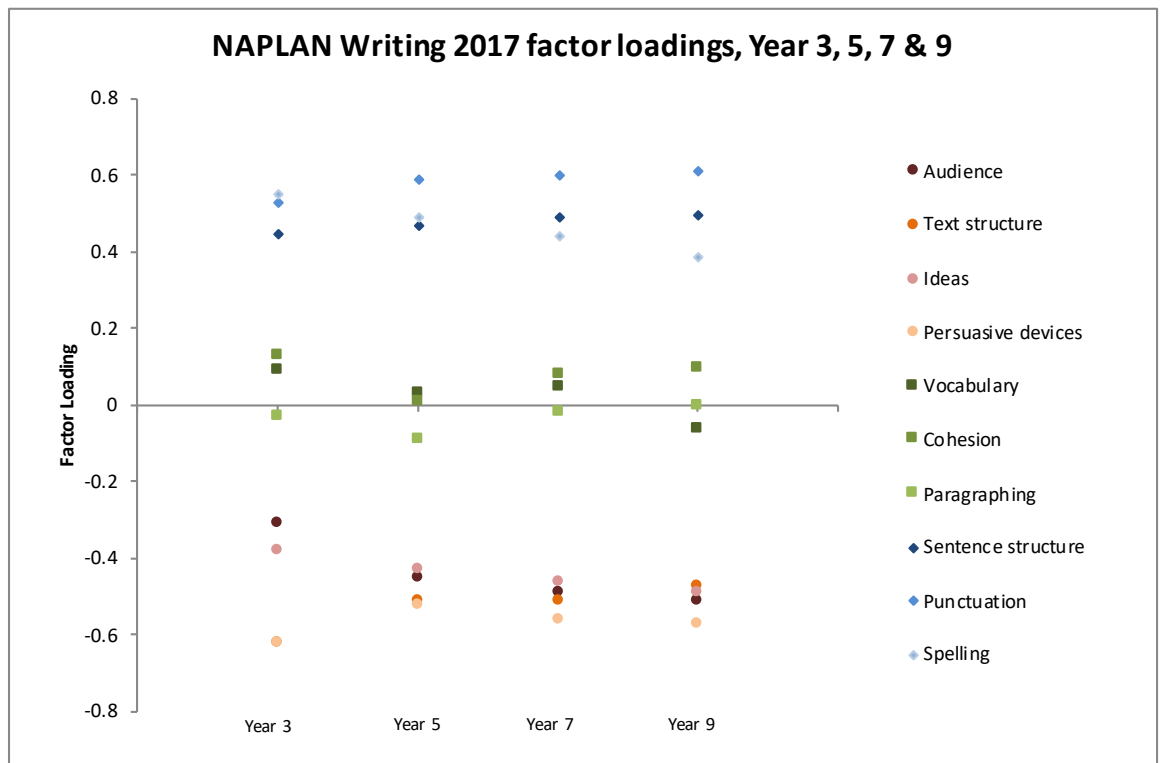
Some traits that include both word and whole text writing features are located midway between them.

In other words, writing test data from NSW government students shows that, while the ten writing traits fit a uni-dimensional model reasonably well, there is evidence that some of our students are:

- stronger (or weaker) on word/sentence level writing traits and/or
- weaker (or stronger) on whole text level writing features.

Figure 1:

Plot of factor loadings from the PCA on the standardised residuals for the NAPLAN 2017 writing tests from Year 3 to Year 9



Moreover, it is worth noting that the performance contrast between these two sets of writing trait scores is consistent with literature on writing and analysis of writing data. For example, researchers (Peters & Smith, 1993) asserted that writing analysis needs to take account of both the **authorial** and **secretarial** aspects of writing, with the authorial aspect encompassing those whole text level writing features such as the organisation of ideas and information to communicate to an audience, and secretarial aspect encompassing the surface or mechanical aspects of writing, such as spelling, grammar and punctuation. Other researchers have argued that many teachers remain focused on the secretarial aspects of writing and neglect the authorial role (Fang & Wang, 2011), and that creating a balance between the authorial and secretarial aspects of writing in teaching is required. It's important that students not only master the skills of how to write correct sentences, but also learn how to write effectively – i.e. how to convey their message and anticipate the needs of the reader, order their thoughts and ideas and carefully choose words and sentences that best convey meaning (Christie, 2005; Wing Jan, 2009).

The fact that the internal structure exhibited in our writing assessment data is consistent with findings from prior research on the writing construct domain provides another piece of evidence supporting the validity of the writing test results, and consequently, the use of writing data in our skillset analysis.

4. Minimum group size required when applying the proposed methodology

The proposed methodology is used to examine a collection of responses, produced by a group of students, on a set of items assessing a common skillset.

A key question then, is what is the minimum number of responses needed for the analyses to produce reliable information? This depends on the minimum sample size required for the resulting distribution of the sum of categorical (or Bernoulli) distributed random variables to approach a normal distribution so statistical significance of the difference between observed sum score and expected sum score can be tested.

To answer this question, six random samples of two, four, six, eight, ten and twelve students were drawn from the 2017 NAPLAN Year 3 writing test data. Three writing traits including Audience (score range 0-6), Paragraphing (0-3) and Punctuation (0-5) were selected for investigation. Figures 2 to 4 show the distribution of the sum of writing trait scores for the selected group of students, i.e. the probabilities of observing each sum score, for different sample sizes, for each of the three selected traits. The kurtosis and skewness for the resulting distribution are also reported in Table 4.

Our investigations suggest a sample of eight students (equivalent to a sample collection of 8 student-item responses) appeared to be sufficient to yield a symmetric and unimodal distribution as shown in Figures 2 to 4. The kurtosis and skewness of the resulting distributions reported in Table 4 also reveal that a sample size of 8 yields a distribution that is close to a normal distribution with kurtosis equal to 3 and skewness equal to 0. In accordance with results from this analysis, all school-level reports (or any group level reports) are suppressed for groups with less than 10 students.

Figure 2:

Resulting distribution of the sum of categorical distributed random variables for 'Audience' for different sample sizes

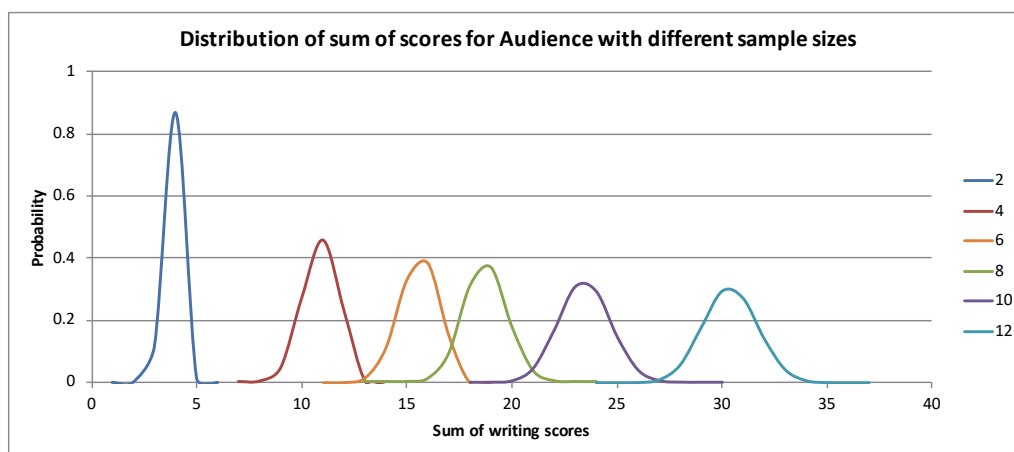
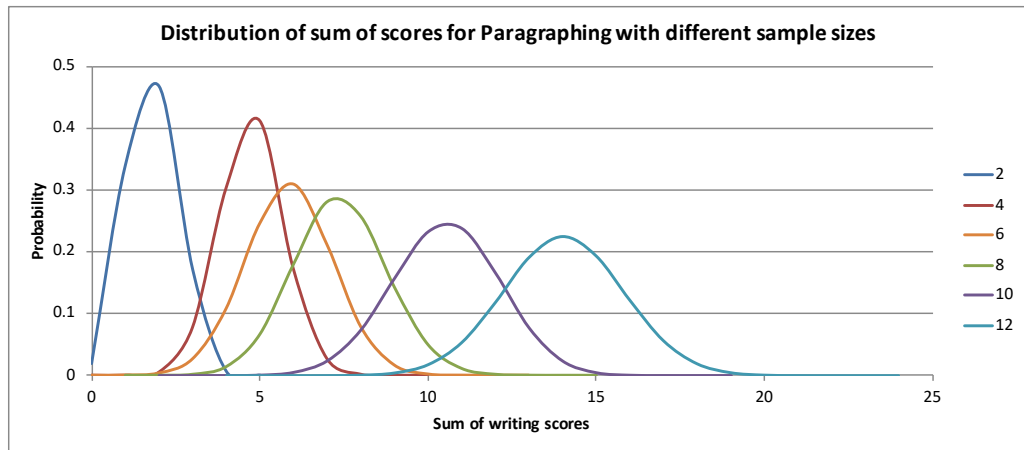
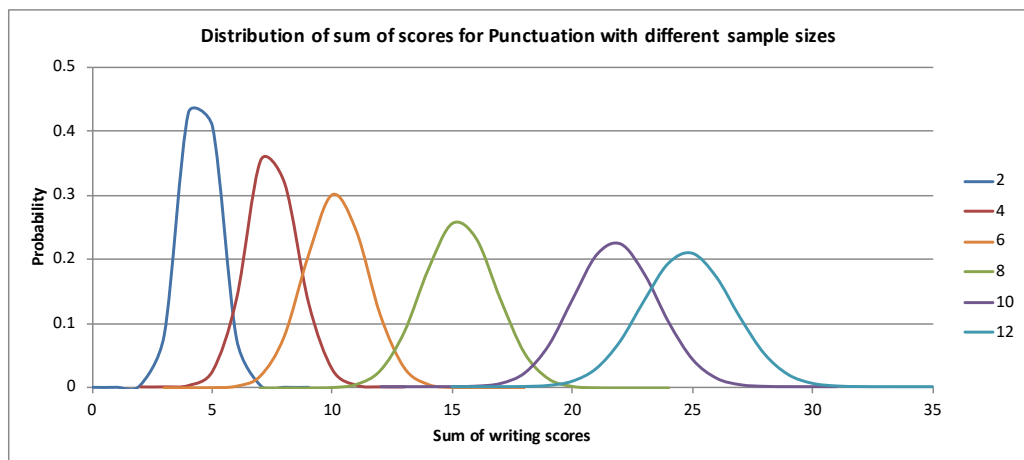


Figure 3:

Resulting distribution of the sum of categorical distributed random variables for 'Paragraphing' for different sample sizes

**Figure 4:**

Resulting distribution of the sum of categorical distributed random variables for 'Punctuation' for different sample sizes

**Table 4:**

Kurtosis and skewness of the resulting distributions for Figures 2 to 4

Writing trait	Sample size	Kurtosis	Skewness
Audience	2	7.257	-1.681
Audience	4	2.617	-0.293
Audience	6	2.784	-0.297
Audience	8	2.985	0.049
Audience	10	2.914	0.058
Audience	12	2.937	0.039
Paragraphing	2	2.545	0.156
Paragraphing	4	3.019	0.046
Paragraphing	6	3.008	-0.003
Paragraphing	8	3.020	0.070
Paragraphing	10	2.961	-0.039
Paragraphing	12	3.004	0.003
Punctuation	2	3.007	0.017
Punctuation	4	3.131	0.072
Punctuation	6	3.046	0.012
Punctuation	8	3.019	0.007
Punctuation	10	2.989	0.036
Punctuation	12	2.999	0.015

5. Model results using NAPLAN writing data (2013 – 2017)

This section presents the results from applying this methodology to NAPLAN writing. For each scholastic year and each calendar year (2013-2017), a Polytomous Rasch model is first fitted, and then the proposed methodology applied, treating each school as a target group. As a result, for each trait, each school in a given year is classified into one of the five categories 'significantly above expectation', 'possibly above expectation', 'as expected', 'possibly below expectation', and 'significantly below expectation'. The categories take into account both statistical and practical significance, as specified in Table 1.

5.1 Proportion of schools identified as weaker or stronger on a particular writing trait in a given year

Using NAPLAN Year 3 2017 writing test as an example, the proposed method classified 1,246 schools with at least 10 NAPLAN participating students into four categories. Table 5 reports the number of schools being classified into each category according to the decision rules included in Table 1 ('as expected' has been excluded from the table as it represents the vast majority of results).

Table 5:

Number of schools in the top and bottom two categories for NAPLAN 2017 Year 3 writing

Marking criteria	Performance category				% of schools in the four categories (excluding 'as expected')
	Significantly above expectation	Possibly above expectation	Possibly below expectation	Significantly below expectation	
Audience	3	29	38	2	5.8%
Text structure	47	107	111	26	23.4%
Ideas	5	40	53	13	8.9%
Persuasive devices	26	78	94	19	17.4%
Vocabulary	0	17	19	1	3.0%
Cohesion	2	18	33	2	4.4%
Paragraphing	111	138	148	38	34.9%
Sentence structure	2	75	70	16	13.1%
Punctuation	36	108	120	25	23.2%
Spelling	45	149	109	61	29.2%

Note that the denominator for the last column is the number of schools with at least 10 NAPLAN participating students.

Table 5 shows that some writing traits such as Paragraphing and Spelling have a higher percentage of schools being classified into the top two and bottom two categories while Audience, Vocabulary and Cohesion have a much lower percentage. These percentages remain fairly stable across calendar years and across scholastic grades. These results are not surprising given the fit statistics for each trait reported in Table 2 where Paragraphing for example showed a larger misfit (underfit) relative to other traits.

5.2 Volatility of school-level model results

For face validity, volatility of school level model results was examined. If there is significant volatility in model results – e.g. there is a large proportion of schools identified as weaker on a particular trait in one year and then ‘stronger’ on the same trait in the next year, it calls into question the face validity of the model results, as it is unlikely that strength or weakness of teaching programs would be drastically different from year to year for a large number of schools. High level volatility in model results also calls into question the utility of the proposed method as the results wouldn’t be useful for planning of teaching programs or intervention.

The writing trait ‘Spelling’ in Year 3 NAPLAN writing test was used to demonstrate the relatively small amount of volatility in the school-level results. For this analysis, any schools with fewer than ten NAPLAN participating students in any one of the adjacent years were excluded from the cross-tabulation. Percentages of schools are reported in Tables 6 to 9 in cross-tab format for two adjacent years from 2013 to 2017. Overall more than 50% of schools remain in the ‘As expected’ category across two adjacent years. Less than 20% of the schools move from the top two or bottom two categories to the ‘As expected’ category in the following year or vice versa. Less than one per cent of schools move from the top category to the bottom category or vice versa.

Table 6:

Cross-tabulation of performance categories for writing trait ‘Spelling’ between 2013 and 2014, NAPLAN Year 3 writing

Year and performance categories		2014				
		Significantly above	Possibly above	As expected	Possibly below	Significantly below
2013	Significantly above	4 (0.3%)	6 (0.5%)	21 (1.8%)	1 (0.1%)	0 (0%)
	Possibly above	7 (0.6%)	28 (2.4%)	67 (5.6%)	7 (0.6%)	0 (0%)
	As expected	19 (1.6%)	72 (6.1%)	705 (59.4%)	79 (6.7%)	30 (2.5%)
	Possibly below	4 (0.3%)	3 (0.3%)	78 (6.6%)	19 (1.6%)	4 (0.3%)
	Significantly below	0 (0%)	1 (0.1%)	21 (1.8%)	7 (0.6%)	3 (0.3%)

Table 7:

Cross-tabulation of performance categories for writing trait ‘Spelling’ between 2014 and 2015, NAPLAN Year 3 writing

Year and performance categories		2015				
		Significantly above	Possibly above	As expected	Possibly below	Significantly below
2014	Significantly above	2 (0.2%)	7 (0.6%)	25 (2.1%)	0 (0%)	0 (0%)
	Possibly above	6 (0.5%)	24 (2%)	71 (5.9%)	7 (0.6%)	1 (0.1%)
	As expected	20 (1.7%)	77 (6.4%)	687 (57.4%)	85 (7.1%)	33 (2.8%)
	Possibly below	1 (0.1%)	5 (0.4%)	92 (7.7%)	13 (1.1%)	5 (0.4%)
	Significantly below	1 (0.1%)	0 (0%)	26 (2.2%)	4 (0.3%)	5 (0.4%)

Table 8:

Cross-tabulation of performance categories for writing trait ‘Spelling’ between 2015 and 2016, NAPLAN Year 3 writing

Year and performance categories		2016				
		Significantly above	Possibly above	As expected	Possibly below	Significantly below
2015	Significantly above	1 (0.1%)	5 (0.4%)	21 (1.8%)	2 (0.2%)	1 (0.1%)
	Possibly above	4 (0.3%)	28 (2.3%)	72 (6%)	7 (0.6%)	2 (0.2%)
	As expected	19 (1.6%)	61 (5.1%)	719 (59.9%)	77 (6.4%)	27 (2.3%)
	Possibly below	0 (0%)	6 (0.5%)	83 (6.9%)	19 (1.6%)	3 (0.3%)
	Significantly below	1 (0.1%)	0 (0%)	29 (2.4%)	9 (0.8%)	4 (0.3%)

Table 9:

Cross-tabulation of performance categories for writing trait ‘Spelling’ between 2016 and 2017, NAPLAN Year 3 writing

Year and performance categories		2017				
		Significantly above	Possibly above	As expected	Possibly below	Significantly below
2016	Significantly above	6 (0.5%)	2 (0.2%)	16 (1.3%)	2 (0.2%)	1 (0.1%)
	Possibly above	10 (0.8%)	23 (1.9%)	56 (4.7%)	10 (0.8%)	1 (0.1%)
	As expected	40 (3.3%)	81 (6.8%)	673 (56.1%)	105 (8.8%)	23 (1.9%)
	Possibly below	2 (0.2%)	3 (0.3%)	73 (6.1%)	26 (2.2%)	12 (1%)
	Significantly below	1 (0.1%)	0 (0%)	24 (2%)	5 (0.4%)	5 (0.4%)

The second piece of analysis performed is to examine the proportion of schools classified into four categories (significantly above, significantly below, possibly above, possibly below) across calendar years, for each grade cohort separately. The aim is to see if these proportions change significantly from one year to the next. Significant variations are indicative of the variability in the model parameters from one year to the next.

Again only results using NAPLAN Year 3 writing data are presented here, but they are similar across grades. We used three writing traits (Audience, Sentence Structure and Spelling) for illustration. Figures 5 to 7 show the percentage of schools in each category for the three selected traits. As shown in these graphs, the year-to-year changes in the percentages of schools identified in each of the categories are less than 1 percentage point for most categories. This pattern is similar across other traits.

Figure 5:

Percentage of schools in each performance category for 'Audience' from 2013 to 2017, NAPLAN Year 3 writing

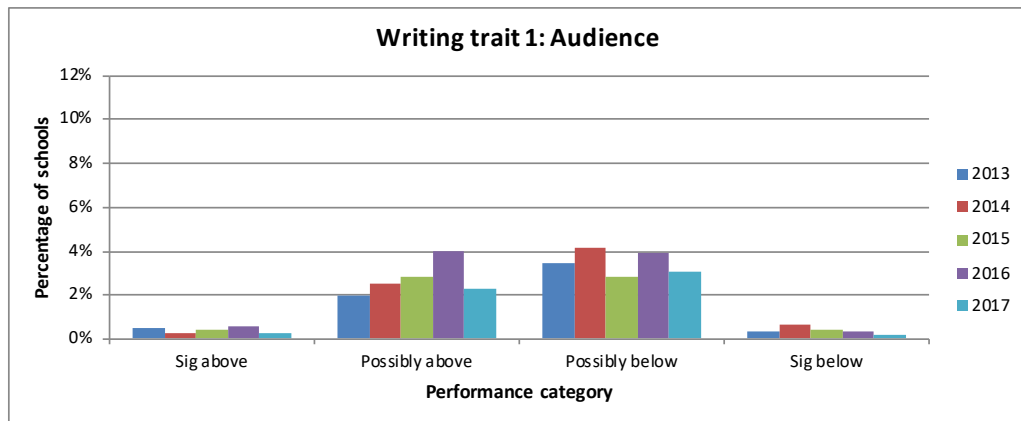


Figure 6:

Percentage of schools in each performance category for 'Sentence structure' from 2013 to 2017, NAPLAN Year 3 writing

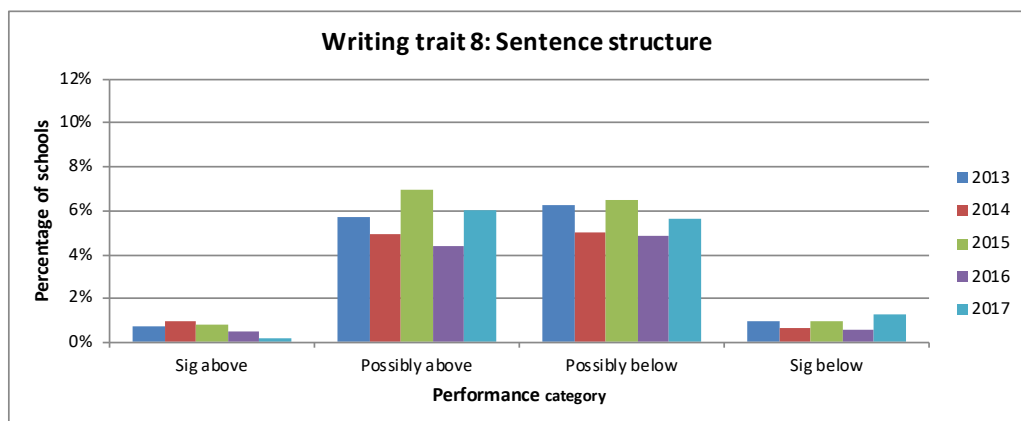
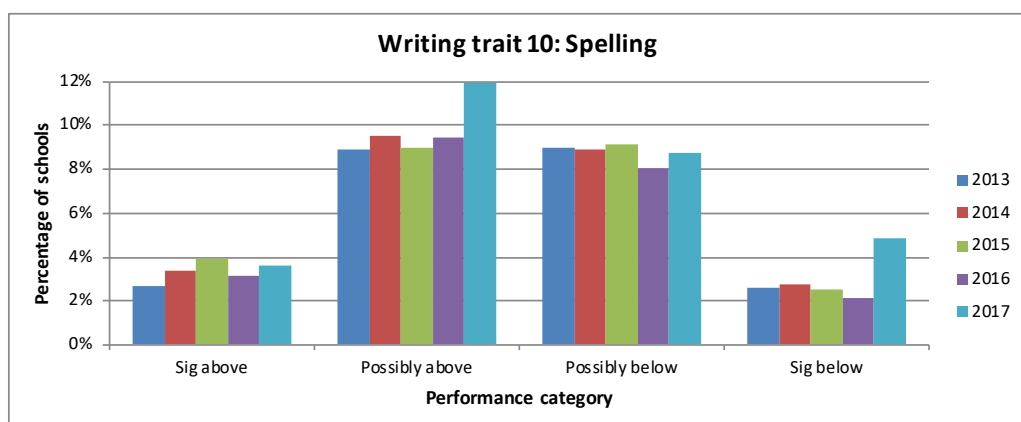


Figure 7:

Percentage of schools in each performance category for 'Spelling' from 2013 to 2017, NAPLAN Year 3 writing



6. Validation of model results

6.1 External validity

To validate the results from the models, we looked at whether model results align with external measures in the way we expected. If they do, this would be a piece of supporting evidence for the validity of the proposed methodology.

To do this, we compared the Spelling trait scores in the NAPLAN writing assessment to the scores from the NAPLAN Spelling test. If our model works as intended, and if it identifies that a Year 3 cohort in a school performed better (or worse) than expected on the Spelling trait (relative to these students' overall writing scores), then this cohort should also be more likely to receive higher (or lower) average Spelling test scores than the average Spelling scores for the Year 3 cohorts in other schools with similar overall writing scores.

For Year 3 and Year 9 separately, Figures 8 and 9 show the relationship between the school average scaled score from NAPLAN Spelling test (Y-axis) against the school average scaled score from NAPLAN writing test (X-axis). Only schools with at least ten NAPLAN participating students are included on the graphs. Each school is marked on the graph, colour coded by the categories our proposed method has classified the schools into for the Spelling trait. Specifically, blue crosses and light blue dots represent schools which were identified through our method as performing better than expected on the Spelling trait with circles used for possibly above and crosses for significantly above expectation. Likewise, red crosses and light red dots represent schools that were identified as performing worse than expected on Spelling.

Starting with Figure 8, we can see a number of patterns of results that suggest that the methodology is working as expected:

1. School average NAPLAN writing scores and Spelling scores are positively correlated, which is not surprising given the two tests assess two aspects of literacy and that spelling is part of the rubric for writing assessment.
2. For schools with the same average writing scores, those identified as performing significantly above expectation (blue crosses) on the Spelling trait are mostly located above the red crosses which are schools identified as performing significantly below expectation. This indicates that schools performing better on the Spelling trait also tended to score higher on the Spelling test.
3. Schools in the possibly below (light red hollow dots) and possibly above expectation (light blue hollow dots) categories are mostly located respectively in the lower and upper middle of the chart.

Similar patterns are replicated in Figure 9, although not as marked as those in Figure 8, due to the smaller number of high schools being classified into the 'significantly above' or 'significantly below' categories.

These patterns provides a sound piece of evidence to support the claim that the methodology is working as intended, and it does produce appropriate and reliable information about the strength and weakness in students' performance.

Figure 8:

Plot of school average NAPLAN spelling score against NAPLAN writing score for Year 3 in 2017

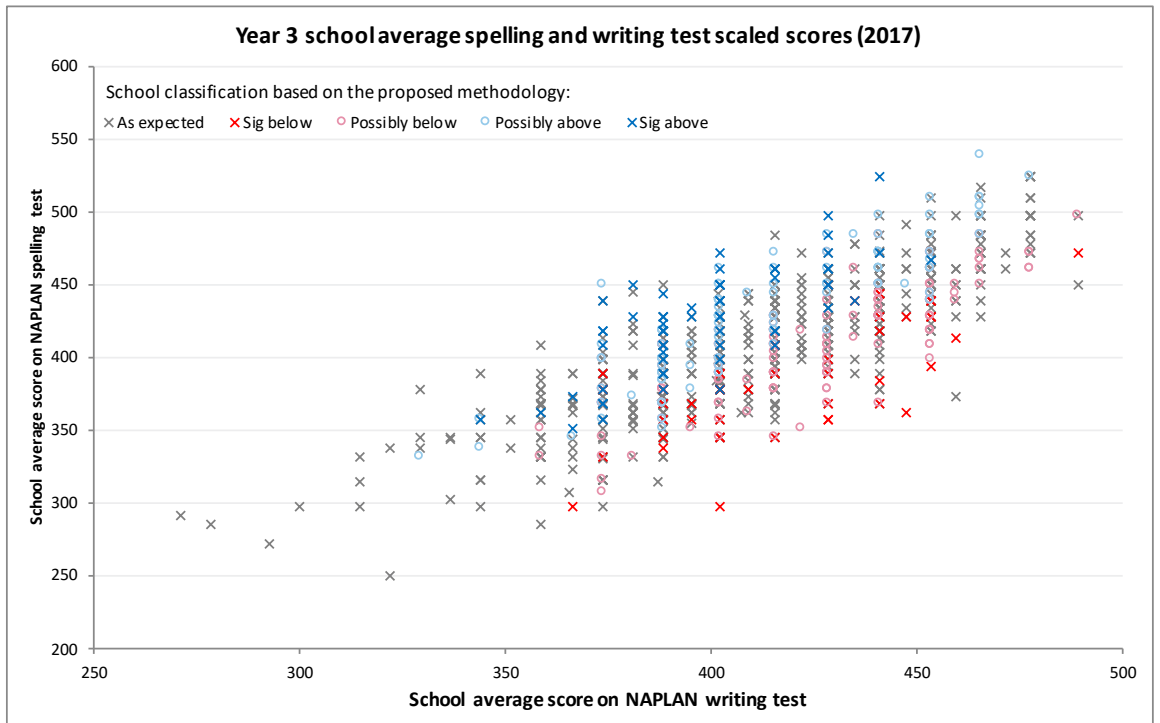
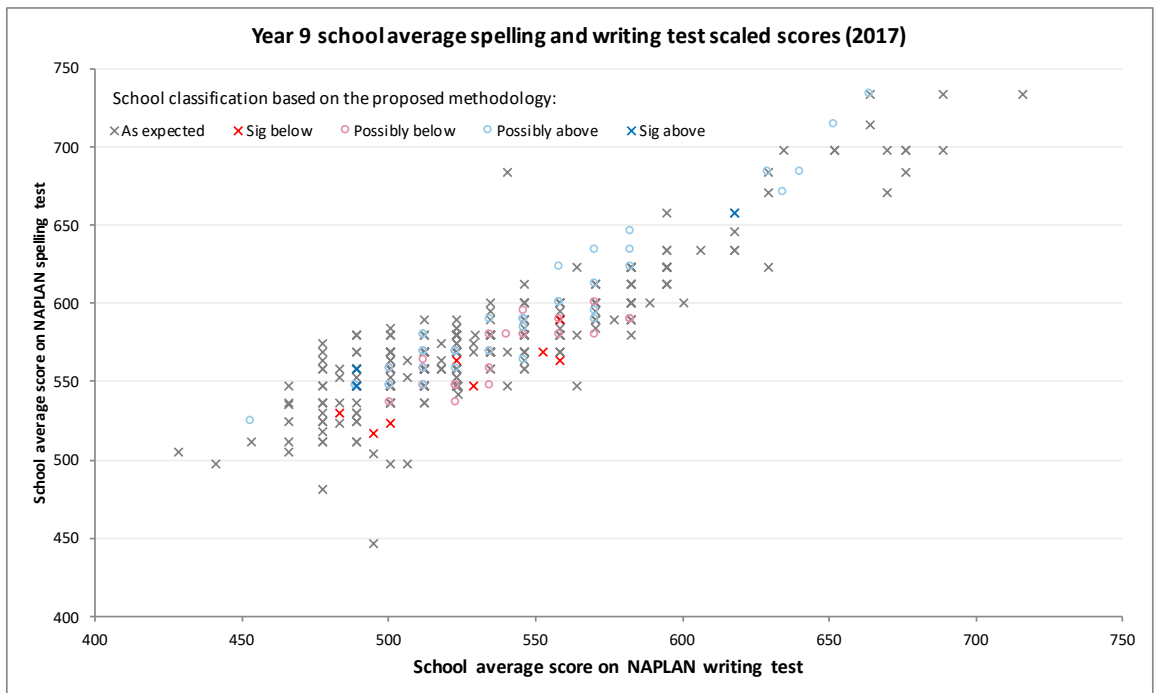


Figure 9:

Plot of school average NAPLAN spelling score against NAPLAN writing score for Year 9 in 2017



7. Reporting options

Reports were developed from an iterative process of design, development and feedback. After a number of workshops with selected groups of teachers, school leaders and curriculum experts, we developed a suite of prototype reports which we thought could be useful to provide to schools and to networks. During these workshops, we guided participants through the proposed methodology and rationale before recording their interpretations of the report meanings, possible school applications and recommendations for changes. Following these workshops, a group of 10 schools trialled the reports to provide further insight into the interpretability and utility of the reports.

7.1 Multiple-school report

The first proposed report is illustrated at Figure 10 which provides NSW Public Schools Directors (or program coordinators) a view of how multiple schools in their networks (or a program) performed on different writing traits.

Figure 10:

Prototype of the proposed school-level report for Directors and program coordinators

School	No. of students	Audience	Text structure	Ideas	Persuasive devices/ Character and setting	Vocabulary	Cohesion	Paragraphing	Sentence structure	Punctuation	Spelling
1	47										
2	58										
3	30										
4	24										
5	14										
6	44										
7	34										

Performance category	Colour label
Significantly above expectation	
Possibly above expectation	
Typical performance	
Possibly below expectation	
Significantly below expectation	

7.2 Individual school trends report

The second type of report includes 5 years of data for an individual school and identifies the writing traits that their students performed stronger or weaker on than expected. This report aims to inform Principals, instructional leaders and teachers about the pattern of strengths and weaknesses of students at a school over a five-year period. The table also contains, for each year, the number of students who participated in the test and the mean writing score for the school. The mean score is provided for context because changes in the pattern of strengths and weaknesses may occur with changes in mean score.

Figure 11:

Prototype table providing trend information on areas of strength or weakness for a grade cohort in a given school, from 2013 to 2017

Genre	Year	Scholastic year	No. of students	Mean scale score	Audience	Text structure	Ideas	Persuasive devices/ Character and setting	Vocabulary	Cohesion	Paragraphing	Sentence structure	Punctuation	Spelling
Persuasive	2013	3	46	425										
Persuasive	2014	3	47	349										
Persuasive	2015	3	39	414										
Narrative	2016	3	35	408										
Persuasive	2017	3	47	401										

7.3 Student level reports

In Section 3.3 we outlined how our writing data from NAPLAN demonstrated that students may acquire different achievement levels for authorial aspects of writing vs secretarial aspects of writing.

Authorial traits are whole text level features that include audience, text structure, ideas and persuasive devices or character and setting traits.

Secretarial traits are word/sentence level features that include sentence structure, punctuation and spelling traits.

Teachers and school leaders from the focus groups indicated that it would be useful to know who these students are and what their patterns of strength and limitations in this regard. Based on those discussions, we developed a sub-domain report using the PCA method outlined in Section 3.3. Figure 12 is a prototype table showing, for a list of students, traits where students performed stronger or weaker than expected. This list (Figure 12) provides information on each individual trait, with the traits grouped into sub-domains. Figure 14 provides a list of students with information on overall performance in each sub-domain.

For a given trait, this list identifies students where the actual score a student received was more than half of a score point above or below the expected trait score, based on the student's overall ability estimate. Bands resulting from the NAPLAN writing test are included to enable comparisons amongst student with similar writing performances.

Figure 12:

Prototype list of students with trait performance information

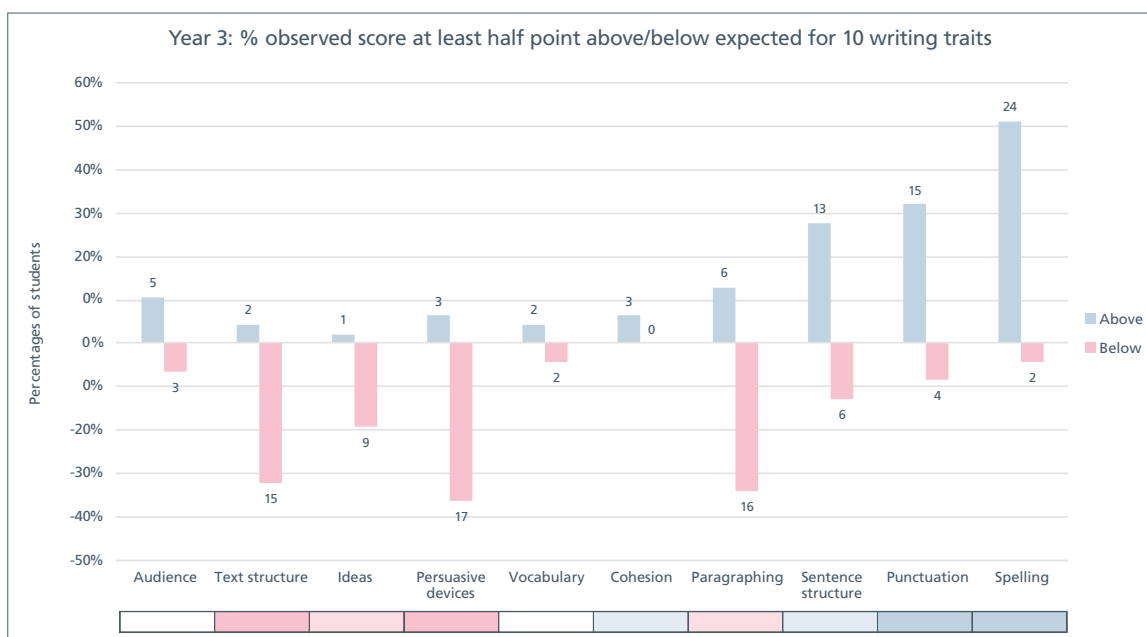
Student	Band	Authorial writing traits				Vocabulary	Cohesion	Paragraphing	Secretarial writing traits		
		Audience	Text structure	Ideas	Persuasive devices				Sentence structure	Punctuation	Spelling
1	2										
2	2										
3	3										
4	3										
5	3										
6	3										
7	3										
8	3										
9	3										
10	3										
11	3										
12	3										
13	3										
14	3										
15	3										
16	3										
17	3										
18	3										
19	4										
20	4										
21	4										
22	4										
23	4										
24	4										
25	4										
26	4										
27	4										
28	4										
29	4										
30	4										

Description	Colour label
Actual trait score is above expectation	
Actual trait score is below expectation	

Figure 13:

Prototype graph summing individual student trait performance information

A graphical representation of the student-level information has also been developed, with the intended audience being the school executives and teachers. The graph summarises the number and proportions of students identified as achieving above or below expected performance for each trait (see Figure 13).



As this report and graph use individual student results, these reports are less reliable, with the impact of inconsistencies in markers' judgements at the trait score level being less likely to be 'ironed out' than in the school level report.

An alternative way to provide student-level information is to group traits into higher level sub-domains and identify those students whose performance on a sub-domain is worth investigating. Figure 14 shows what such a sub-domain report could look like.

Figure 14:

Prototype list of students with performance information on authorial/secretarial writing traits

Student	Band	Authorial (Audience, Text structure, Ideas, Persuasive devices, Characters and setting)	Secretarial (Sentence structure, Punctuation, Spelling)
1	2		
2	2		
3	3		
4	3		
5	3		
6	3		
7	3		
8	3		
9	3		
10	3		
11	3		
12	3		
13	3		
14	3		
15	3		
16	3		
17	3		
18	3		
19	4		
20	4		
21	4		
22	4		
23	4		
24	4		
25	4		
26	4		
27	4		
28	4		
29	4		
30	4		

7.4 Trial outcomes

A survey of teachers who had trialled the reports suggested teachers generally found the proposed reports useful. However, the survey results also indicated instances of inconsistent interpretations of the reports, partly due to a tendency of users to confuse relative strengths and weaknesses with absolute student scores in each skillset. Qualitative analysis of the participating teacher survey results showed that many participants, while feeling confident in their interpretations, were not correctly conceptualising the reports.

This result is unsurprising as this new methodology represents a significant shift in teacher expectations and understandings of the NAPLAN data, particularly given the established practice of relying on the proportion of correct answers for a test item to gauge students' strengths and weaknesses.

We will continue to explore whether these interpretation challenges are able to be remedied using additional support material and to support the effective use of these resources.

8. Conclusions and next steps

The move to NAPLAN online has significant potential benefits in improving student test engagement, faster feedback on results and more tailored testing. The increased complexity of reporting when fewer students are exposed to the same items, however, creates a challenge for how to provide clear, accessible and reliable information to inform teaching programs.

This paper has outlined a method of analysing the test results in response to this challenge. It represents a shift in the focus of test analysis from an individual item to a learning area or a skillset that is commonly assessed by a group of items. While providing valuable and statistically robust information, trialling with schools has identified that further work is needed for the reports to be correctly understood by school staff.

There is a clear need for a comprehensive support package to communicate this alternate methodology. With this in mind, in 2019, CESE (in collaboration with other parts of the department) will undertake a project with a small selection of targeted schools to develop and trial resources that support understanding and analysis of reports, and to investigate and document potential implications of these reports for teaching and learning.

Writing researchers have identified a need for the development and validation of 'integrated writing assessment systems that provide immediate instructionally relevant multi-vector data to teachers so that they are better equipped for pinpointing writing problems and responding accordingly' (Troia, 2007). In this regard, we hope the analysis and reports proposed, which are intended to provide diagnostic and actionable information about students' strengths and weaknesses in their writing to teachers and school leaders, would help meet this need.

References

- Australian Curriculum Assessment and Reporting Authority (ACARA) 2016, *Online assessment*, viewed 7 January 2019, <<https://www.acara.edu.au/assessment/online-assessment>>.
- Australian Curriculum Assessment and Reporting Authority (ACARA) 2018, *National literacy learning progression*, viewed 30 June 2018a, <<https://australiancurriculum.edu.au/media/3673/national-literacy-learning-progression.pdf>>.
- Australian Curriculum Assessment and Reporting Authority (ACARA) 2018, *National numeracy learning progression*, viewed 30 June 2018b, <<https://www.australiancurriculum.edu.au/media/3806/numeracy-learning-progression.pdf>>.
- Board of Studies NSW 2012a, *English K-10 syllabus*, viewed 30 June 2018, <<https://educationstandards.nsw.edu.au/wps/portal/nesa/k-10/learning-areas/english-year-10/english-k-10>>.
- Board of Studies NSW 2012b, *Mathematics K-10 syllabus*, viewed 30 June 2018, <<https://educationstandards.nsw.edu.au/wps/portal/nesa/k-10/learning-areas/mathematics/mathematics-k-10>>.
- Christensen, K & Kreiner, S 2013, 'Item fit statistics', in K Christensen, S Kreiner & M Mesbah (eds), *Rasch models in health*, John Wiley & Sons, Hoboken, pp. 83-102.
- Christie, F 2005, *Language education in the primary years*, University of New South Wales Press, Sydney.
- Fang, Z & Wang, Z 2011, 'Beyond rubrics: Using functional language analysis to evaluate student writing', *Australian Journal of Language and Literacy*, vol. 34, no.2, pp. 147-165.
- Hambleton, R & Swaminathan, H 1985, *Item response theory: Principles and applications*, Kluwer, Boston.
- Linacre, J 1998, 'Detecting multidimensionality: Which residuals data-type works best?', *Journal of Outcome Measurement*, vol. 2, no.3, pp. 266-283.
- Linacre, J 2002, 'Optimizing rating scale category effectiveness', *Journal of Applied Measurement*, vol. 3, no.1, pp. 85-106.
- Linacre, J 2009, *A user's guide to WINSTEPS*, Chicago: winsteps.com.
- Loevinger, J 1957, 'Objective tests as instruments of psychological theory', *Psychological Reports*, vol. 3, no. 3, pp. 635-694.
- Messick, S 1996, 'Validity and washback in language testing', *Educational Testing Service*, vol. 13, no. 1, pp. 1-18.
- Myford, C & Wolfe, E 2003, 'Detecting and measuring rater effects using many-facet Rasch measurement: Part I', *Journal of Applied Measurement*, vol.4, no.4, pp. 386-422.
- Peters, M & Smith, B 1993, *Spelling in context: Strategies for teachers and learners*, NFER-Nelson, London.
- Rasch, G 1960/1980, *Probabilistic models for some intelligence and attainment tests*, Expanded edn, The University of Chicago Press, Chicago.
- Troia, G 2007, 'Research in writing instruction: What we know and what we need to know', in M Pressley, A Billman, K Perry & K Reffitt (eds), *Shaping literacy achievement*, The Guilford Press, New York, pp. 129-156.
- Wing Jan, L 2009, *Write ways: Modelling writing forms*, 3rd edn, Oxford University Press, South Melbourne.
- Wright, B & Linacre, J 1994, 'Reasonable mean-square fit values', *Rasch Measurement Transactions*, vol. 8, no.3, pp. 370-371.
- Wright, B & Stone, M 1999, *Measurement essentials*, 2nd edn, Wide Range, Wilmington, Delaware.
- Yildirim, H, Yildirim, S & Verhelst, N 2014, 'Profile analysis as a generalized differential item functioning analysis method', *Education and Science*, vol. 39, no. 172, pp. 49-64.

Appendix A

Marking rubric of NAPLAN writing test

NAPLAN writing test – narrative genre

Writing trait	Description of narrative writing marking criterion
Audience	The writer's capacity to orient, engage and affect the reader
Text structure	The organisation of narrative features including orientation, complication and resolution into an appropriate and effective text structure
Ideas	The creation, selection and crafting of ideas for a narrative
Character and setting	Character: The portrayal and development of character Setting: The development of a sense of place, time and atmosphere
Vocabulary	The range and precision of contextually appropriate language choices
Cohesion	The control of multiple threads and relationships across the text, achieved through the use of grammatical elements (referring words, text connectives, conjunctions) and lexical elements (substitutions, repetitions, word associations)
Paragraphing	The segmenting of text into paragraphs that assists the reader to negotiate the narrative (Note: Different number of categories compared to persuasive writing marking rubric)
Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences
Punctuation	The use of correct and appropriate punctuation to aid the reading of the text
Spelling	The accuracy of spelling and the difficulty of the words used

NAPLAN writing test – persuasive genre

Writing trait	Description of persuasive writing marking criterion
Audience	The writer's capacity to orient, engage and persuade the reader
Text structure	The organisation of the structural components of a persuasive text (introduction, body and conclusion) into an appropriate and effective text structure
Ideas	The selection, relevance and elaboration of ideas for a persuasive argument
Persuasive devices	The use of a range of persuasive devices to enhance the writer's position and persuade the reader
Vocabulary	The range and precision of contextually appropriate language choices
Cohesion	The control of multiple threads and relationships across the text, achieved through the use of grammatical elements (referring words, text connectives, conjunctions) and lexical elements (substitutions, repetitions, word associations)
Paragraphing	The segmenting of text into paragraphs that assists the reader to negotiate the narrative (Note: Different number of categories compared to narrative writing marking rubric)
Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences
Punctuation	The use of correct and appropriate punctuation to aid the reading of the text
Spelling	The accuracy of spelling and the difficulty of the words used


Appendix B

Item fit statistics – Infit and Outfit

Denote the observed response by X_{ni} and the total number of students by N , two item-fit statistics, namely the infit and outfit, can then be derived from a comparison between the expected and observed responses to investigate how well the data fits the model (Wright & Stone, 1999). Omitting observations with extreme scores, that is, students who scored zero or scored maximum for all items, the infit and outfit can be derived as:

$$\text{Infit} = \frac{\sum_{n=1}^N (X_{ni} - E_{ni})^2}{\sum_{n=1}^N V_{ni}} \quad \text{and} \quad \text{Outfit} = \frac{\sum_{n=1}^N [(X_{ni} - E_{ni})^2 / V_{ni}]}{N}.$$

Infit is an information-weighted fit statistic which is more sensitive to unexpected behaviour affecting responses to items near the student's latent ability. Outfit is an outlier-sensitive fit statistic, more sensitive to unexpected behaviour by students on items far from the student's latent ability. Both mean-square statistics have an expected value of 1.0, and a range from 0 to positive infinity. Values less than 1.0 indicate over-fit; that is, data is too predictable with respect to model expectations, causing summary statistics such as reliability indices, to report inflated results. Values greater than 1.0 indicate under fit; that is, there is more un-modelled noise in the data than expected. High mean-squares are considered a much greater threat to the validity than low mean-square values, because they suggest a possible violation of the uni-dimensionality requirement (Linacre, 1998; Linacre, 2002; Myford & Wolfe, 2003).

A network diagram with various sized nodes and connecting lines, set against a teal background.

Author: Dr Lucy Lu and Dr Wai-Yin Wan

Centre for Education Statistics and Evaluation
GPO Box 33, Sydney NSW 2001, Australia

Visit our website to subscribe to the CESE newsletter

☎ 7814 1527

🗨 Yammer

✉ info@cese.nsw.gov.au

🌐 cese.nsw.gov.au

NSW Department of Education April 2019

Please cite this publication as:

Centre for Education Statistics and Evaluation (2019), Identifying potential strength and weakness in key learning areas using data from NAPLAN tests, NSW Department of Education www.cese.nsw.gov.au