Review of the NAPLAN marking rubrics for Narrative and Persuasive Writing:

Summary of Evidence and Recommendations

Dr Stephen Humphry,
University of Western Australia

August, 2020

Contents

Executive summary

This report reviews the National Assessment Program Literacy and Numeracy (NAPLAN) marking rubrics for Narrative and Persuasive writing assessment. The first part of the report comprises a summary of evidence and recommendations, identifying areas of suggested rubric and marker training changes. The second part of the report comprises further detailed data analysis and results. There is substantial agreement between inferences for Persuasive and Narrative, though there are also points of difference.

Key observations and recommendations include:

- Evidence suggests that markers commonly use scoring patterns across criteria, which tend to produce superfluous information for the purposes of differentiating among levels of writing capability. A recommendation is made to reduce the criteria where there is evidence of superfluous information due to pronounced rating tendencies. It is also recommended that where any new criteria are created that combine criteria, markers are given explicit discretion to make professional on-balance type judgements based on features in score categories.
- Criteria with score categories that are not sufficiently differentiated are identified, and recommendations made that these criteria are reduced, and frequency-based rules are avoided. It is also recommended that in Spelling and Punctuation, there is an implicit understanding that students with higher level performances are given the highest possible score rather than being marked down for a small number of errors.
- Due to severity of rating tendencies for high-performing students, it is recommended that holistic score criteria are used for such performances on the authorial criteria (Audience, Text Structure, Ideas, Persuasive Devices and Vocabulary).

1. Rating tendencies and superfluous information

1.1 Background

The term *rating tendency* is used here to refer to markers' use of common scoring patterns across criteria. As outlined to follow, there is evidence of pronounced rating tendencies that reflect and/or produce superfluous information for the purpose of differentiating among levels of writing capability. The pronounced rating tendencies may occur for various reasons. Perhaps the most likely reason is that global or holistic perceptions influence scores on multiple criteria intended to be distinct. Central tendency effects may or may not also occur. However, the data analysis conducted for this report do not definitively indicate the cause, they only point to observations that are useful for practical decisions about changes to the NAPLAN marking guide.

It is stressed that rating tendencies *per se* are not the issue. There must be rating tendencies across criteria for assessments to be reliable. The key question is whether there are pronounced rating tendencies of a kind that reflect and/or result in superfluous information.

1.2 Evidence

The evidence indicates some level of superfluousness when individual markers assess the following clusters:

- Persuasive: Ideas, Audience, Text Structure, Persuasive Devices, and Cohesion;
- Narrative: Ideas, Audience, Text Structure, and Cohesion.

Spelling and vocabulary, as a criterion pair, also have stronger within-marker rating tendencies than occur between-marker, indicating some level of superfluousness. Data analysis indicate some level of superfluousness for other criteria also, as detailed further in other sections below.

Evidence that there is superfluous information comes from multiple sources. First, in full NAPLAN data analyses, subsets of criteria show indicators of statistical dependence (violations of independence). Second, within-marker rating tendencies are stronger than between-marker tendencies, in certain cases much stronger. Third, triangulation with pairwise comparison scale locations, using residuals correlations, provides evidence that is generally consistent with the first two sources of information.

It seems likely that rating tendencies that reflect and/or produce superfluous information in NAPLAN marking are at least partly a result of cognitive load and the nature of the guide, therefore not purely marker behaviour that is easily rectified. Considerable research would likely be needed to definitively establish the underlying factors at play.

The different sources of evidence do not generally indicate strong within-marker rating tendencies producing superfluous information for Paragraphing, Punctuation, and Spelling, although there are some exceptions. Also, separate issues in relation to these criteria are observed below.

1.3 Recommendation 1

Reduce the number of criteria in sets for which there is evidence of superfluous information, based upon: (a) objectives for assessment within the programme; and (b) empirical reference to performances that have been ordered based on pairwise comparisons. The strength of rating tendencies indicate use of one criterion to capture most, or even all, of those things described within existing authorial choices criteria.

Reference to empirical data is desirable so that newly formed criteria describe features that experienced markers perceive as tending to occur together in actual writing performances. There are pre-existing Persuasive and Narrative performances that have been ordered both overall and with respect to so-called authorial choices criteria, which could be used as a basis for identifying and or validating proposed new criteria.

1.4 Recommendation 2

It is recommended that in any newly formed criteria that encompass multiple existing criteria, markers are given explicit discretion to make professional judgement based on features described in score categories. Put in the opposite way, it is not recommended that there are strict requirements, expressly or implied, that all features described in a score category must be evident to award the score.

The rationale for this recommendation is that should markers feel it necessary to be convinced all criteria implied by descriptors have been met, this is likely to lead to conservatively low scores. Students can have strengths and weaknesses in development that can reasonably be considered to balance in terms of the overall quality of a performance. Pairwise comparison exercises have involved such on-balance judgement across aspects, and resulting scale locations correlate very highly with current rubric scores (notwithstanding observations made to follow in this report).

2. Criteria with score categories that are not sufficiently differentiated

There are two criteria for which certain score categories do not appear sufficiently clearly delineated for effective use by markers to differentiate levels of writing performance. Triangulation of criterion scores with the locations of scripts based on pairwise comparisons indicates that there is substantial overlap in the estimated locations of students who obtain adjacent scores. The details are as follows:

- For Paragraphing there appears to be relatively poor delineation between score categories 2 and 3.
- For Punctuation there appears to be relatively poor delineation between all score categories, particularly 3, 4 and 5 (noting very few students are awarded 5).

The data analysis used to identify rating tendencies that produce superfluous information indicates that to a reasonable extent, judgements of Paragraphing scores are influenced by judgements of Audience scores, *but not the other way around*. This statement is mainly applicable to score categories 2 and 3.

Consistent with the points above, the threshold discrimination is relatively low (poor) for Punctuation score categories 3 and 4. Lower discrimination is not taken to indicate an issue in itself, but it is relevant given evidence from pairwise comparison data. Threshold discrimination appears typical for Paragraphing, though this may be due to the influence of Audience.

Spelling categories appear to be generally delineated other than for the highest category (in which there are few data).

Vocabulary score category 1 covers a very large range of abilities based on pairwise comparisons, with it being awarded to students that overlap in ability greatly with score category 2 as well as 0. This suggests that it is unlikely the three categories can be used to effectively differentiate levels of writing capability.

As noted above, Spelling and Vocabulary have stronger within-marker rating tendencies than occur between-marker, which indicates some superfluous information. The effect appears mutual, rather than scores on one criterion being influenced by scores on another criterion. In addition, the rating tendencies for Spelling and Vocabulary appear to be more pronounced for Persuasive performances than for Narrative performances, based on the trial data available for analysis.

Research conducted by Humphry and Heldsinger (2019) also shows that for a group of experienced NAPLAN markers, Punctuation, Spelling and Paragraphing as defined in the NAPLAN marking guide, were seldom deemed essential to making direct comparisons between writing performances of moderately high ability and above (approximately average Year 7 and higher). Further, these criteria were deemed relevant quite infrequently, across the whole range of performances, relative to the other criteria. See Figure 6.

2.1 Recommendation 1

It is recommended that the number of conventions criteria are reduced. The data do not clearly indicate which criteria to retain. It is recommended that multiple distinct components are only included in score categories if this is based on empirical evidence. It is recommended that frequency-based rules are avoided in score categories.

2.2 Recommendation 2

Whatever choices are made with respect to inclusion of criteria relating to Spelling and Punctuation, it is recommended that the application of score categories are limited to students in Year 3 with low capabilities through to, at most, average Year 7 students. In this case, most or all students producing higher level performances would be given the highest possible score for these criteria. Markers would then use other criteria to differentiate between higher levels of writing capability.

3. Issues specific to higher score categories

Several sources of information indicate greater dependence among the highest score categories on some criteria. This may operate in combination with lack of delineation of the score categories for certain criteria, if markers reinforce common score patterns on criteria lacking delineation.

The sources of evidence are as follow. First, referenced to pairwise locations, there is a greater range of rubric scores for a given range of pairwise locations in the higher ability range then below this (see Figure 1). Second, for a number of criteria there is greater reinforcement of common score patterns within-marker than between-marker in the higher score categories. Third, the variance of logits obtained from analysis of the rubric is significantly higher in Year 9 than in Years 3 and 5. A substantial proportion of Year 9 students obtain scores above 30.

The scatterplot of rubric scores against pairwise locations in Figure 1 shows an increased slope in the upper region, approximately average Year 7 and higher. For a given range of logits on the pairwise axis, there is a greater range of rubric scores above a score of about 30. Pairwise locations were referenced to individual criterion scores for 2013 NAPLAN data, showing that the highest score categories tend to overlap with the next highest categories for most of the criteria.

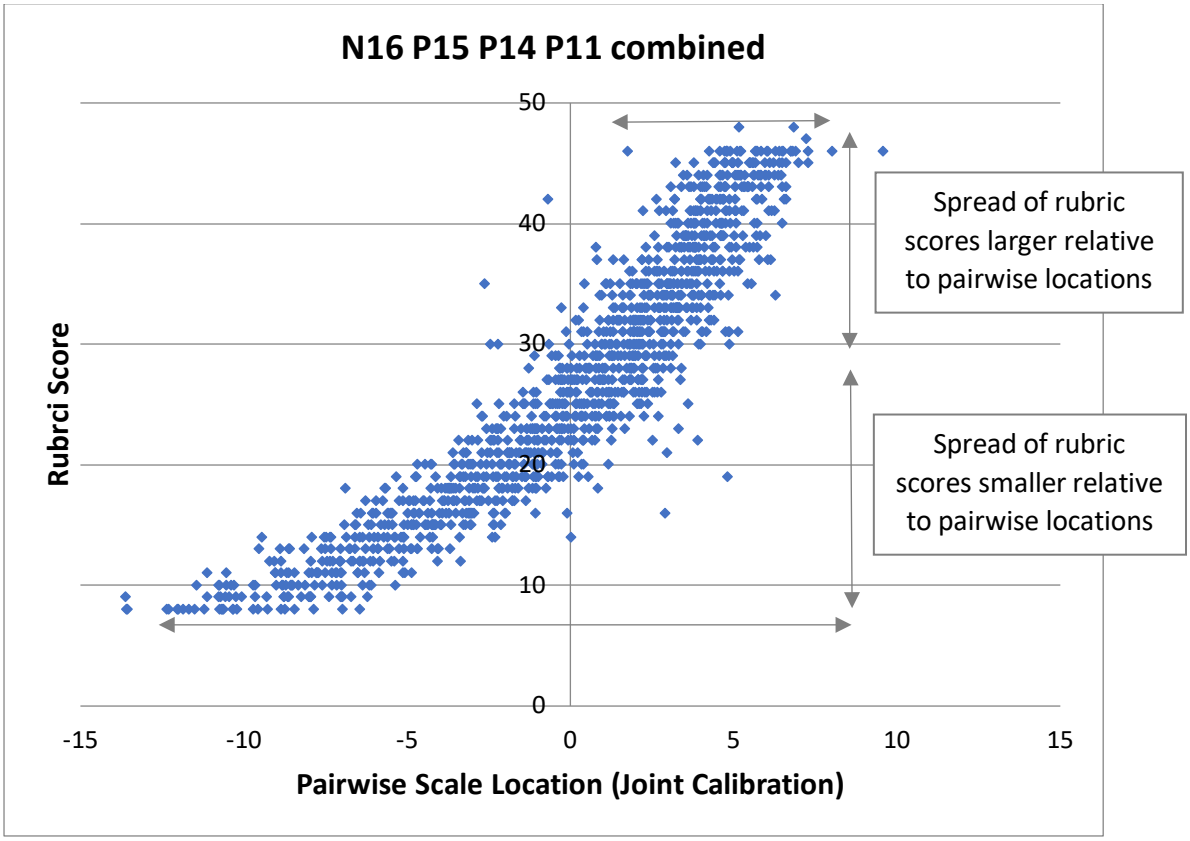

Figure 1. Scatterplot of rubric scores against pairwise locations across multiple NAPLAN years for Narrative and Persuasive Writing

In combination with the far larger spread of abilities in Year 9, this suggests rating tendencies produce especially superfluous information in the higher regions.

There is a modest correlation between pairwise scale location and Outfit mean square (r = 0.26), which suggests the change in slope in Figure 1 is due partly to smaller separation of performances in the upper range, on the scale based on pairwise comparisons (x-axis).

At the same time, however, there is also evidence that the increase in slope is due in large part to stronger rating tendencies in the higher score categories. Within-marker rating tendencies are much stronger than between-marker rating tendencies in higher score categories for some criterion pairs. Some *illustrative* cases for double-marked Persuasive trial data are shown in Table 1. The ratio is the number of times the score pattern occurred when the same marker assessed the criteria to the number of times the score pattern occurred when different markers assessed the criteria.

Table 1. Illustrative cases of within-marker reinforcement of common score patterns in high score categories for NAPLAN Persuasive Writing

Criterion pair	Score pattern	Ratio (within to between)
Audience, Ideas	5,4	1.4
Audience, Ideas	6,5	1.9
Audience, Text Structure	6,4	1.6
Audience, Punctuation	6,4	1.7
Audience, Punctuation	5,5	1.8
Audience, Spelling	6,6	1.7
Text Structure, Ideas	6,6	1.4
Vocab, Spelling	0,0	1.4
Vocab, Spelling	5,6	2.0

For example, the Audience, Punctuation score pattern {6,4} occurred 1.7 times as often within-marker than between-marker, which is a great deal more. This reinforcement of scoring patterns is likely to be a significant reason that the slope in Figure 1 is very steep in the top part. Reinforcement occurs across a set of criteria means that when a marker gives a performance a higher score on one of those criteria, this is reinforced on multiple criteria, increasing the total score in an exaggerated fashion relative to a marker who gives a performance a lower total score on relevant criteria. Some of the high scores are superfluous.

As stated above, also consistent with this tendency, the spread in logits is far greater for Year 9 students than students in Years 3 and 5. The Year 9 standard deviation is on average 25% larger than the Year 5 standard deviation (ratio of standard deviations range from 1.22 to 1.34) between 2013 and 2019. There is no similar tendency for Reading or other assessments.

Table 2. Historical standard deviations for NAPLAN Narrative and Persuasive Writing by year level

Year Level	Calendar Year (prompt)						
	2013 (Pers)	2014 (Pers)	2015 (Pers)	2016 (Narr)	2017 (Narr)	2018 (Pers)	2019 (Narr)
3	70.5	71.4	67.4	62.0	65.7	69.5	63.1
5	70.1	69.7	66.1	63.0	64.4	67.8	64.5
7	76.2	75.4	76.0	70.6	76.0	74.0	71.0
9	87.1	85.3	85.2	77.0	86.2	83.3	78.7

Although it is possible that variance actually increases for Year 9 students, it is unlikely to be a coincidence that the slope in Figure 1 increases at a point above which many Year 9 students are located; it is more likely that the increase in the spread is an artefact of increased rating tendencies that are associated with superfluous information.

If needed to accomplish objectives, qualitative evidence could also be examined to further test whether the increased spread is an artefact. In particular, it is possible to select performances with rubric scores above 30 that have: (i) similar locations based on pairwise comparisons; but (ii) substantially different rubric scores. If such qualitative examination suggests the performances are indeed similar on-balance, this would confirm that score differences are exaggerated due to rating tendencies associated with superfluous information. Ideally, such confirmation would be done in a manner whereby expert markers have no knowledge of the pairwise locations. Preliminary examinations of precisely this kind were conducted during the NAPLAN data analysis period in 2013 when issues were identified, and these suggested that score differences were artificially exaggerated. However, only a small number of performances were examined.

The reason data from pairwise comparisons and double-marked performances are useful is that analysis of data using an IRT model on its own has a self-limiting capacity to detect dependencies. Dependencies expand the range of logit estimates (person location and item threshold/delta) in a way that tends to make the common score patterns more expected. In turn, this limits the degree to which dependencies will be detected through residual correlations.

3.1 Recommendation

Given the severity of rating tendencies in high score categories for authorial choices criteria, irrespective of how many authorial choices are included, it is recommended that holistic score criteria are used for the most capable performances.

Further Detail of Data Analysis and Results

4. Background for NAPLAN Writing assessments

4.1 Assessment tasks for NAPLAN Persuasive Writing

The assessment tasks are designed to elicit persuasive writing from students. Students are presented with a stimulus to which they need to respond and a different stimulus is used for Years 3 / 5 and Years 7 / 9.

4.2 Assessment criteria for NAPLAN Persuasive Writing

NAPLAN Persuasive Writing performances are assessed with respect to 10 criteria, each of which have a different number of categories.

Criterion		Definition	Number of categories
1.	Audience	The writer's capacity to orient, engage and persuade the reader	0 - 6
2.	Text structure	The organisation of the structural components of a persuasive text (introduction, body and conclusion) into an appropriate and effective text structure	0 - 4
3.	Ideas	The selection, relevance and elaboration of ideas for a persuasive argument	0 - 5
4.	Persuasive devices	The use of a range of persuasive devices to enhance the writer's position and persuade the reader	0 - 4
5.	Vocabulary	The range and precision of contextually appropriate language choices	0 - 5
6.	Cohesion	The control of multiple threads and relationships across the test, achieved through the use of referring words, ellipses, text connectives, substitutions and word associations	0 - 4
7.	Paragraphing	The segmenting of the text into paragraphs that assists the reader to follow the line of argument	0 - 3
8.	Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences	0 - 6
9.	Punctuation	The use of correct and appropriate punctuation to aid the reading of the text	0 - 5
10.	Spelling	The accuracy of spelling and the difficulty of the words used	0 - 6

The criteria can be loosely categorised as **authorial choices** which include Audience, Text Structure, Ideas, Persuasive Devices and Vocabulary; and **conventions** which include Paragraphing, Sentence Structure, Punctuation and Spelling

4.3 The structure of the marking guide for NAPLAN Persuasive Writing

The guide comprises of the 10 assessment criteria listed above, performance descriptors, and exemplars.

For each criterion, there are a number of categories or score points; and a performance descriptor is provided for each of the categories. Additional information is also provided to explain terminology used in the descriptor and/or to provide other information about the descriptor.

Approximately 19 exemplars of varying ability are provided in the guide, to represents the typical range of abilities from Year 3 to Year 9. Each exemplar is a marked performance with scores on each of the assessment criteria. An annotation is provided with each exemplar explaining how the exemplar was marked, and in particular the reason for the awarded scores.

5. Rating tendencies

The analysis of double-marked trial scripts indicates rating tendencies that are stronger when the same marker scores criteria than when different markers score criteria. That is, certain patterns of scores across criteria occur more frequently when the same marker assesses the criteria than when different markers assess the criteria.

The amplification of rating tendencies occurs for some pairs of criteria, but not other pairs, as detailed below. Generally, reinforced rating tendencies are evident for pairs of authorial choices criteria, and especially pairs involving the first three criteria.

Rating tendencies can occur because the alignment between the criteria is implied by the structure of a rubric (Humphry & Heldsinger, 2014). However, they could also occur for other reasons such as cognitive load in conjunction with raters' familiarity with common scoring patterns. If there are too many closely related criteria and/or if there is too much cognitive load:

> "[J]udges may have little choice but to make spurious distinctions either by defaulting to a pattern of common scoring (akin to a response set) or through recourse to a global judgment" (Humphry & Heldsinger, 2014, p. 256).

The examples in Table 1 illustrate rating tendencies. For example, the score pattern Audience 5, Ideas 4 occurred 1.4 times as often when the same marker assessed both criteria as when different markers assessed Audience and Ideas. On the other hand, the score pattern Audience 3, Ideas 3 occurred 1.17 times more often when marked by the same marker than different markers (not shown). This is still noticeably more, albeit not as pronounced a difference.

It might be said this indicates higher and artificial 'consistency' for within-marker scoring than between-marker scoring. However, this description may be taken to imply that markers alone cause stronger rating tendencies, when there are actually other factors at play. The key question is whether objective evidence indicates that the rating tendencies reflect and/or result in superfluous information.

It seems likely that, at least some of the time, amplification of rating tendencies occurs due to a combination of cognitive load and time pressure. The cognitive load placed on markers is quite high when looking at multiple criteria, each with multiple score categories, and with reference to exemplars. Given time pressures, markers may default to an overall judgement

about the quality of a performance and they may tend to rely on known common score patterns.

In some cases, there may be some overlap in the meaning of criteria such that there is literally some redundancy. However, separate analysis was conducted to try to identify likely candidates for overlap, based on highly frequent score patterns, and this did not detect a great deal of clear overlap. This seems to indicate that clear overlap is not a significant factor, rather it is only part of the picture. Humphry and Heldsinger (2014) describe a combination of criteria overlap and recourse to global judgements as likely contributors to strong rating tendencies.

5.1 Persuasive—within-marker versus between-marker evidence

Table 3 shows, for each pair of criteria, the ratio of the spread of thresholds/deltas marked by: (a) the same maker; vs (b) different markers. Where the ratio is high, this suggests greater within-marker amplification of common score patterns, for example where score patterns such as 5,4 or 2,2 across the criteria occur more frequently when the same marker assesses both criteria than when different markers assess the criteria. The specific common score patterns that are amplified depend on the criteria.

As mentioned above, the issue is not that certain score patterns are more common—that is to be expected due to genuine associations in performances between features of writing described in the criteria. The issue highlighted is that certain common score patterns become *even more common* when the same marker assesses a given pair of criteria than when different markers assess the criteria.

This is unlikely to be due to central tendency alone because it is a relative matter of amplification of score patterns when the same marker assesses certain criterion pairs. However, central tendency may also be occurring and this discussion ought not be taken to suggest that central tendency is not a factor.

Table 3. Ratios of standard deviations of thresholds for criterion pairs for NAPLAN Persuasive Writing

	Aud	TxtStr	Ideas	PersDev	Vocab	Coh	Para	SentStr	Punc	Sp
Aud		1.67	2.09	1.57	1.29	1.38	1.12	1.12	1.07	1.25
TxtStr	1.54		1.85	1.53	1.29	1.23	1.20	1.13	1.07	1.08
Ideas	2.33	1.70		1.59	1.48	1.43	1.10	1.14	1.09	1.17
PersDev	1.74	1.52	1.57		1.34	1.59	1.16	1.11	1.08	1.10
Vocab	1.18	1.15	1.44	1.13		1.33	1.02	1.17	1.11	1.41
Coh	1.45	1.21	1.44	1.48	1.51		1.08	1.30	1.05	1.15
Para	1.26	1.25	1.23	1.22	1.24	1.15		1.17	1.09	
SentStr	1.22	1.46	1.26	1.48	1.34	1.51	1.37		1.17	1.29
Punc	1.08	1.07	1.09	1.10	1.05	1.06	1.09	1.15		1.13
Sp	1.32	1.15	1.35	1.22	1.45	1.34		1.37	1.21	

Each ratio in Table 3 is the ratio of the standard deviation of the thresholds/deltas obtained from (a) within-marker data for the criterion pair relative to (b) between-marker data for the criterion pair. Ratios above 1.4 are highlighted in the table and all other tables showing the same kinds of ratios that follow.

The standard deviations of the thresholds/deltas are shown in Appendix 1. The values in the table are the ratio of the standard deviation of the thresholds/deltas for the criterion listed in the row, for each pair of criteria. For example, the ratio of the standard deviation of the thresholds/deltas for Text Structure, paired with Audience, is 1.54. The ratio for Audience, paired with Text Structure, is 1.67. Generally, these ratios mirror each other in magnitude, though not always.

A higher ratio indicates a tendency toward greater amplification of some score patterns in that pair. Higher ratios give an indication of amplification of some of the common score patterns, although they do not indicate which score patterns are reinforced. Cross-tabulations of scores for pairs of criteria can be used to obtain this information should it be considered useful to examine at this level of detail.

As mentioned in the summary, the data analysis indicates that to a reasonable extent, judgements of Paragraphing scores are influenced by judgements of Audience, *but not the other way around*. The reason is that Paragraphing thresholds are spread out 26% more for within-marker assessments in Audience-Paragraphing pairs (Para row); on the other hand, Audience thresholds are only spread out 12% more for within-marker assessments of the same pair (Aud row). Thus, the effect is not symmetric.

In a similar vein, Spelling thresholds are more spread for within-marker scoring than the other criterion when paired with: Audience, Ideas, Cohesion, and Sentence Structure. The results are similar for Narrative and Persuasive, though the magnitudes of ratios differ somewhat.

However, the thresholds of Audience, Text Structure, Ideas, and Persuasive Devices are not usually more spread within-marker than between-marker in combination with Paragraphing, Punctuation, Sentence Structure and Spelling (the right four columns in Tables 3 and 4 tend to be ratios nearer to 1). This indicates that it is unlikely scores on Audience, Text Structure, Ideas and Persuasive Devices are strongly influenced by scores on the last four criteria. This may be due to the nature of the criteria, though it may also be due at least partly to the ordering of the criteria in the guide.

Spelling and Vocab thresholds are mutually more spread for within-marker scoring than between-marker, indicating mutual artificial consistency.

The analysis indicates that there are amplified rating tendencies among the following criteria: Audience, Text Structure, Ideas and Persuasive Devices. There are also amplified rating tendencies between Cohesion and this main set of four criteria, though less pronounced for the Cohesion-Text Structure pair.

5.2 Narrative—within-marker versus between-marker evidence

Again, each ratio in Table 4 is the ratio of the standard deviation of thresholds/deltas obtained from (a) within-marker data for the criterion pair relative to (b) between-marker data for the criterion pair. The standard deviations of the thresholds/deltas are shown in Appendix 2.

The notable amplifications of rating tendencies are discussed in the summary.

Table 4. Ratios of standard deviations of thresholds for criterion pairs for NAPLAN Narrative Writing

	Aud	TxtStr	Ideas	CharSet	Vocab	Coh	Para	SentStr	Punc	Sp
Aud		1.32	2.08	1.28	1.35	1.10	0.93	1.10	1.00	1.14
TxtStr	1.45		1.74	1.25	1.15	1.13	1.21	1.11	1.05	1.11
Ideas	2.73	1.62		1.30	1.82	1.18	1.15	1.15	1.07	1.13
CharSet	1.80	1.39	1.34		1.15	1.09	1.08	1.13	1.01	1.06
Vocab	1.45	1.24	1.68	1.05		1.54	1.28	0.94	0.95	1.21
Coh	1.33	1.17	1.44	1.44	1.78		1.10	1.45	1.06	1.07
Para	1.24	1.32	1.22	1.18	1.18	1.22		1.08	1.06	1.07
SentStr	1.14	1.65	1.59	1.43	1.16	1.57	0.95		1.14	1.18
Punc	1.10	1.04	1.09	1.00	1.04	1.05	0.95	1.28		1.24
Sp	1.23	1.12	1.21	1.12	1.28	1.17	1.35	1.20	1.14	

5.3 Rating tendencies between markers

The above discussion focused on evidence that rating tendencies are exaggerated when the same marker assesses a given pair of criteria. If this exaggeration occurs due to the imputation of a global judgement, it is possible that rating tendencies between markers also reflect or produce superfluous information. The effect may simply tend to be more pronounced when the same marker assesses both criteria.

In a relevant study, Humphry and Heldsinger (2014) found that with individual criterion marking, there was still evidence of rating tendencies. The fact that within-marker rating tendencies across sets of criteria are stronger than between-marker rating tendencies does not mean that there are no between-marker tendencies due to the nature of the criteria and the cognitive load. Different markers may still form a similar overall judgement about the work some of the time, to some extent.

6. Distinctions in the rubric that are not substantiated in paired comparisons

This section contains evidence for the statements regarding poor delineation between score categories, based on triangulation with pairwise scale locations from the NAPLAN Persuasive Writing pairwise exercise conducted in 2013.

In Figures 2 through 5 to follow, the x-axis is the scale location from on balance pairwise comparisons of performances.

As indicated, for Paragraphing there appears to be relatively poor delineation between score categories 2 and 3. Figure 2 shows that there is effectively complete overlap between the pairwise locations (x-axis) of students who obtain scores of 2 and 3 for Paragraphing (y-axis). Based on this external evidence, the higher score reflects only a very marginal average increase of overall writing capability. The evidence noted earlier suggests that the scores are influenced by Audience scores, and this is likely to be the reason that standard IRT modelling does not show the same issue (it is disguised by the dependence based on within-marker scores).

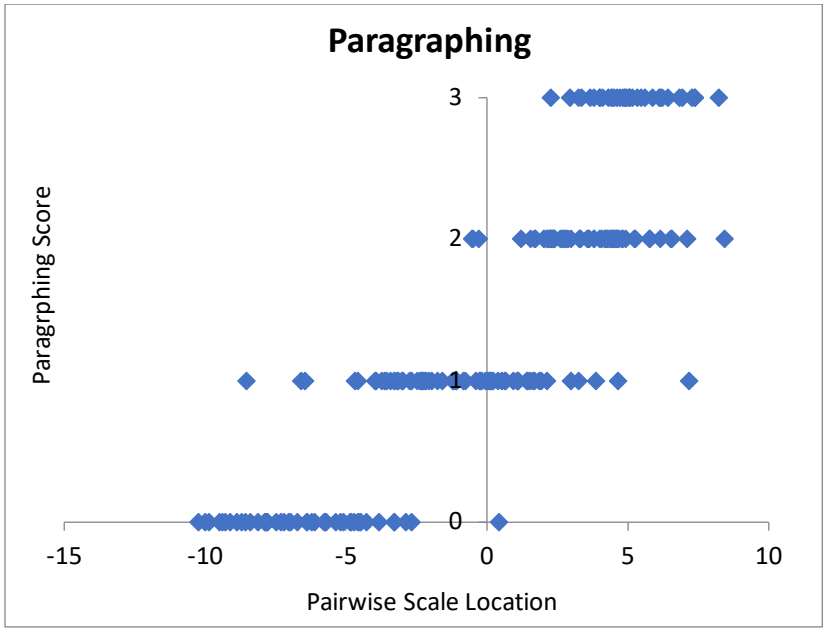

Figure 2. Relation of Paragraphing rubric score to pairwise location based on 2013 NAPLAN Persuasive Writing data

For Punctuation there appears to be relatively poor delineation between all score categories, particularly 3, 4 and 5 (noting very few students are awarded 5). Figure 3 shows substantial overlap of pairwise locations associated with different score categories.

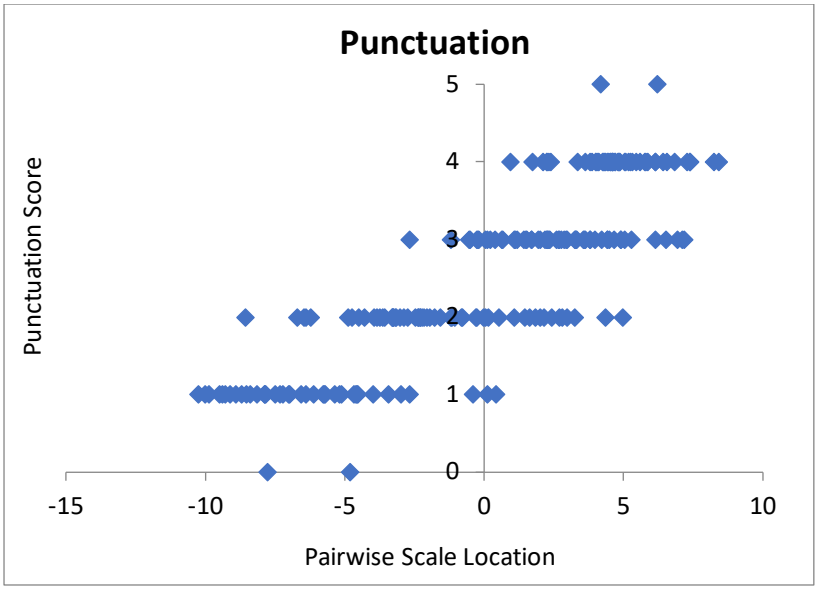

Figure 3. Relation of Punctuation rubric score to pairwise location based on 2013 NAPLAN Persuasive Writing data

Spelling categories appear to be generally delineated other than for the highest category (in which there are few data), as shown in Figure 4.

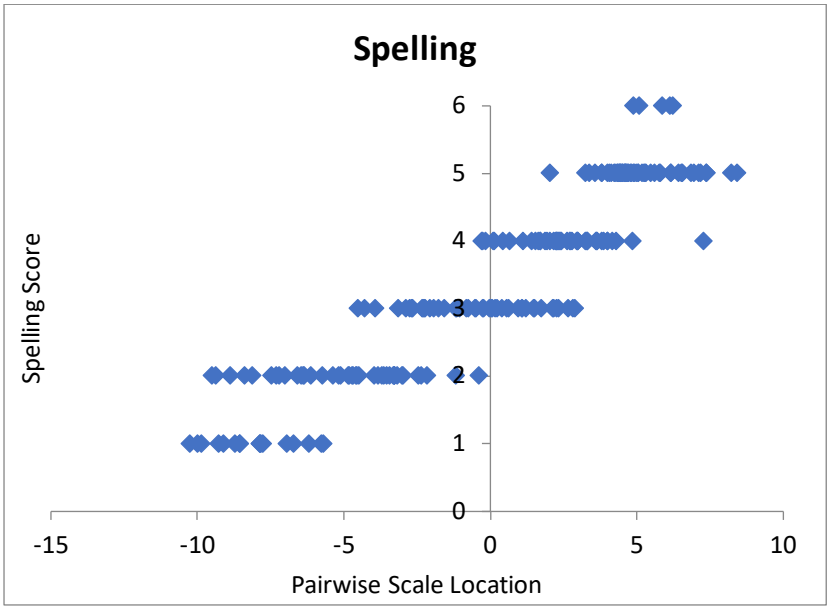

Figure 4. Relation of Spelling rubric score to pairwise location based on 2013 NAPLAN Persuasive Writing data

Figure 5 shows that Vocabulary score category 2 covers a very large range of abilities based on pairwise comparisons, with it being awarded to students that overlap in ability greatly with score category 1 as well as 3. This suggests that it is unlikely the three categories can be used to effectively differentiate levels of writing capability.

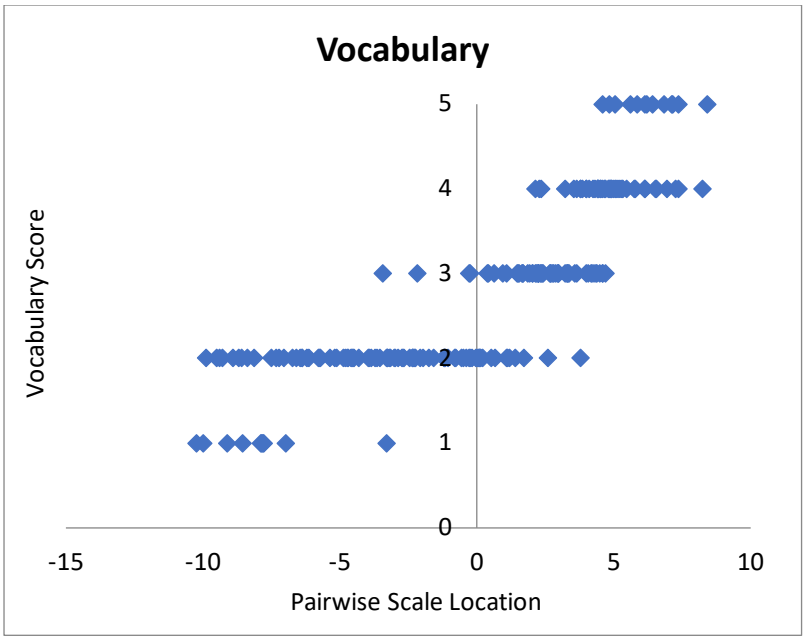

Figure 5. Relation of Vocabulary rubric score to pairwise location based on 2013 NAPLAN Persuasive Writing data

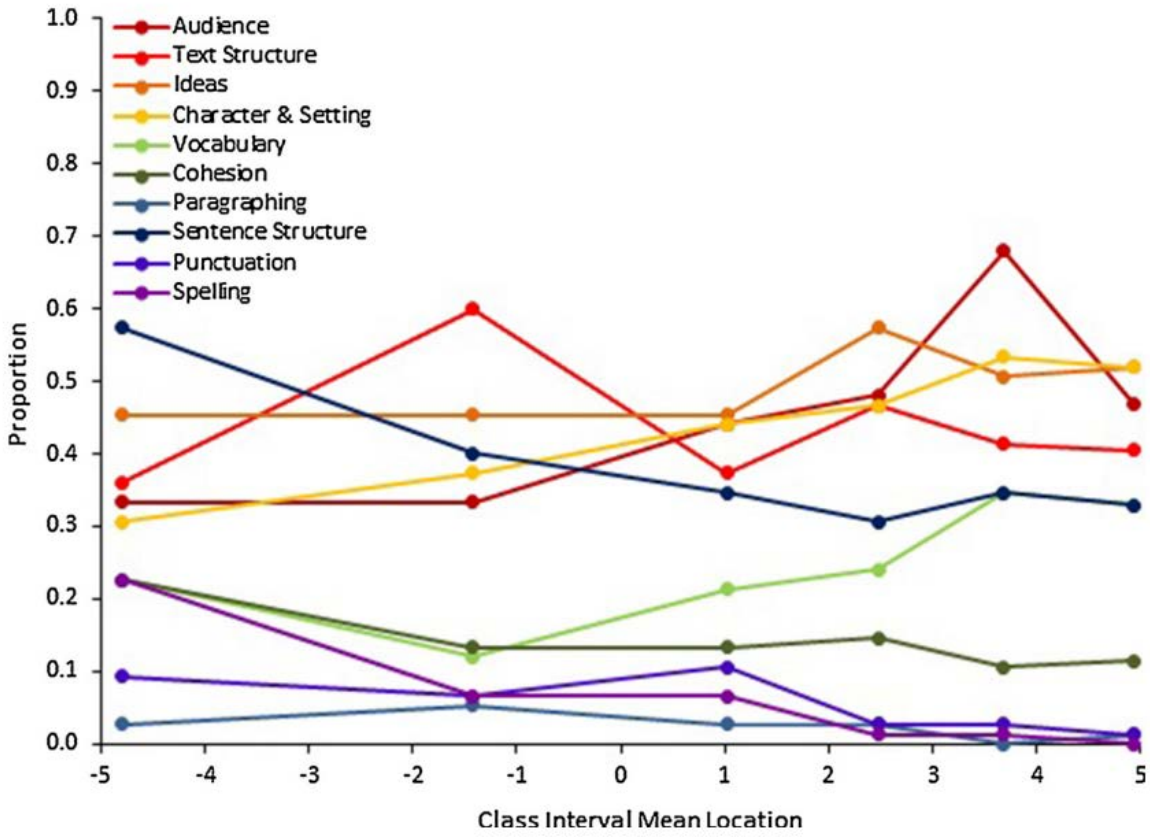

Figure 6. Perceived relevance of criteria for NAPLAN Narrative Writing

As noted in the summary, Research conducted by Humphry and Heldsinger (2019) showed that a group of experienced NAPLAN markers deemed Punctuation, Spelling and Paragraphing as defined in the NAPLAN marking guide, seldom essential to making direct comparisons between writing performances of moderately high ability and above (approximately average Year 7 and higher). Figure 6 summarises the information. Pairs of relatively similar performances were compared, with the x-axis showing the mean scale location for any given pair based on pairwise comparisons. The assessors were asked to indicate, for each pair, which criteria in the NAPLAN guide were regarded as essential to the determination. For further information, please refer to the article.

7. Qualitative Observation

For certain categories in the marking guide there are quite specific, and distinct, components that may or may not tend to occur together in real performances. An example is provided in Table 5. It is not necessarily the case that in performances for which "all sentence punctuation [is] correct", the performances "[provide] accurate markers to enable smooth and efficient reading". It is a qualitative empirical question whether these features occur together in actual performances to an extent that warrants including the very specific component (all sentence punctuation correct) in conjunction with the other component. Issues may arise due to the specificity of the wording if, for example, there are performances that provide accurate markers to enable smooth and efficient reading in which there are one or two cases of incorrect sentence punctuation. This observation is relatively minor in comparison with those outlined above.

Table 5. Example of criterion in which score categories have distinct components

Score	Sample of description
4	"sentence level punctuation mostly correct …"
	"provides adequate markers to assist reading"
5	"all sentence punctuation correct"
	"provides accurate markers to enable smooth and efficient reading"

References

Humphry, S. M. & Heldsinger, S. A. (2014). Common Structural Design Features of Rubrics May Represent a Threat to Validity. *Educational Researcher*, *43*(5), 253-263.

Humphry, S. M. & Heldsinger, S. A. (2019). Raters' perceptions of assessment criteria relevance. *Assessing Writing*, *41*, 1-13. https://doi.org/10.1016/j.asw.2019.04.002

Appendix 1

Table A1. Standard deviations of within- and between-marker thresholds for criterion pairs for NAPLAN Persuasive Writing

	Within-marker									
	Aud	**TxtStr**	**Ideas**	**PersDev**	**Vocab**	**Coh**	**Para**	**SentStr**	**Punc**	**Sp**
Aud		12.90	18.74	13.22	11.72	12.04	11.19	8.30	6.61	10.52
TxtStr	8.73		9.59	8.77	5.82	6.07	6.26	4.59	3.81	4.68
Ideas	18.55	11.13		11.31	10.19	10.37	7.91	7.13	5.51	8.20
PersDev	10.81	8.22	8.91		6.65	7.74	5.75	5.01	3.91	5.04
Vocab	11.62	10.73	15.95	11.46		14.54	14.27	11.71	10.09	14.70
Coh	12.59	7.89	10.94	9.72	12.00		9.71	10.74	6.66	9.44
Para	6.28	5.88	6.12	5.31	5.28	5.78		4.49	3.51	
SentStr	8.64	7.86	8.19	8.50	9.46	11.36	8.88		7.05	8.84
Punc	5.15	4.41	4.89	4.73	4.99	5.60	5.04	5.66		5.21
Sp	9.67	5.64	9.18	6.42	10.89	9.72		8.61	6.41	

	Between-marker									
	Aud	**TxtStr**	**Ideas**	**PersDev**	**Vocab**	**Coh**	**Para**	**SentStr**	**Punc**	**Sp**
Aud		7.72	8.98	8.43	9.05	8.72	9.97	7.41	6.18	8.39
TxtStr	5.67		5.19	5.74	4.51	4.92	5.21	4.07	3.55	4.34
Ideas	7.96	6.53		7.11	6.88	7.28	7.16	6.26	5.07	7.00
PersDev	6.21	5.40	5.66		4.97	4.87	4.95	4.53	3.61	4.57
Vocab	9.84	9.35	11.05	10.15		10.93	13.98	10.02	9.11	10.44
Coh	8.67	6.50	7.62	6.56	7.96		9.01	8.23	6.37	8.21
Para	5.01	4.69	4.96	4.35	4.24	5.05		3.84	3.21	
SentStr	7.10	5.40	6.49	5.73	7.06	7.54	6.47		6.01	6.86
Punc	4.75	4.12	4.51	4.30	4.75	5.27	4.63	4.90		4.61
Sp	7.31	4.90	6.78	5.28	7.49	7.25		6.30	5.29	

Appendix 2

Table A2. Standard deviations of within- and between-marker thresholds for criterion pairs for NAPLAN Narrative Writing

	Within-marker									
	Aud	**TxtStr**	**Ideas**	**CharSet**	**Vocab**	**Coh**	**Para**	**SentStr**	**Punc**	**Sp**
Aud		11.04	23.07	13.74	12.79	10.22	8.75	9.07	7.02	11.03
TxtStr	8.01		10.10	6.98	6.14	5.57	5.74	5.24	4.12	5.29
Ideas	21.90	10.51		9.96	13.35	7.65	5.47	6.89	4.61	7.31
CharSet	12.29	7.21	8.90		7.91	6.13	5.39	7.51	4.69	6.47
Vocab	13.25	10.12	17.26	9.56		11.96	12.04	9.32	8.32	13.38
Coh	9.88	7.11	11.02	9.65	13.17		7.67	14.70	6.28	8.42
Para	5.40	4.82	4.58	4.50	4.58	4.07		4.07	3.70	4.03
SentStr	8.43	9.22	10.75	9.98	9.06	14.13	8.11		6.28	7.80
Punc	4.86	4.24	4.65	4.53	4.91	5.06	4.81	5.45		5.94
Sp	8.26	5.85	8.12	7.11	10.36	7.62	7.69	7.30	5.66	

	Between-marker									
	Aud	**TxtStr**	**Ideas**	**CharSet**	**Vocab**	**Coh**	**Para**	**SentStr**	**Punc**	**Sp**
Aud		8.35	11.07	10.69	9.45	9.32	9.43	8.28	7.01	9.64
TxtStr	5.54		5.79	5.56	5.34	4.93	4.74	4.70	3.92	4.78
Ideas	8.04	6.51		7.66	7.35	6.46	4.76	5.98	4.30	6.48
CharSet	6.82	5.20	6.66		6.89	5.60	5.00	6.63	4.64	6.11
Vocab	9.11	8.18	10.29	9.11		7.79	9.41	9.89	8.79	11.10
Coh	7.43	6.06	7.63	6.72	7.41		6.96	10.17	5.90	7.90
Para	4.37	3.64	3.76	3.82	3.88	3.34		3.76	3.48	3.78
SentStr	7.40	5.59	6.77	6.98	7.79	9.00	8.50		5.50	6.62
Punc	4.42	4.10	4.28	4.55	4.71	4.80	5.08	4.25		4.80
Sp	6.69	5.24	6.74	6.33	8.10	6.49	5.69	6.10	4.97	

Appendix 3 – marking guides and rubrics

Narrative writing marking guide: https://www.nap.edu.au/_resources/2010_Marking_Guide.pdf

Persuasive writing marking guide: https://www.nap.edu.au/_resources/2012_Marking_Guide.pdf

Narrative writing rubric

Criterion	Definition	Number of categories
1. Audience	The writer's capacity to orient, engage and affect the reader	0 - 6
2. Text structure	The organisation of narrative features including orientation, complication and resolution into an appropriate and effective text structure	0 - 4
3. Ideas	The creation, selection and crafting of ideas for a narrative	0 - 5
4. Character and setting	Character: The portrayal and development of character Setting: The development of a sense of place, time and atmosphere	0 - 4
5. Vocabulary	The range and precision of language choices	0 - 5
6. Cohesion	The control of multiple threads and relationships across the text, achieved through the use of referring words, substitutions, word associations and text connectives	0 - 4
7. Paragraphing	The segmenting of text into paragraphs that assists the reader to negotiate the narrative	0 - 2
8. Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences	0 - 6
9. Punctuation	The use of correct and appropriate punctuation to aid the reading of the text	0 - 5
10. Spelling	The accuracy of spelling and the difficulty of the words used	0 - 6

Persuasive writing rubric

Criterion	Definition	Number of categories
1. Audience	The writer's capacity to orient, engage and persuade the reader	0 - 6
2. Text structure	The organisation of the structural components of a persuasive text (introduction, body and conclusion) into an appropriate and effective text structure	0 - 4
3. Ideas	The selection, relevance and elaboration of ideas for a persuasive argument	0 - 5
4. Persuasive devices	The use of a range of persuasive devices to enhance the writer's position and persuade the reader	0 - 4
5. Vocabulary	The range and precision of contextually appropriate language choices	0 - 5
6. Cohesion	The control of multiple threads and relationships across the test, achieved through the use of referring words, ellipses, text connectives, substitutions and word associations	0 - 4
7. Paragraphing	The segmenting of the text into paragraphs that assists the reader to follow the line of argument	0 - 3
8. Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences	0 - 6
9. Punctuation	The use of correct and appropriate punctuation to aid the reading of the text	0 - 5
10. Spelling	The accuracy of spelling and the difficulty of the words used	0 - 6