Improving Learning

WAR: NAPLAN Writing Rubric Review

Final Report

Australian Curriculum, Assessment and Reporting Authority

6 July 2020

ACER

COMPANY INFORMATION

COMPANY	The Australian Council for Educational Research Limited (ACER)
ABN	19 004 398 145
COMPANY ADDRESS	19 Prospect Hill Road Camberwell, Victoria 3124 Australia
WEBSITE	www.acer.org
TELEPHONE	+61 3 9277 5555
FAX	+61 3 9277 5500
CONTACT PERSON	Dr Catherine McClellan
EMAIL ADDRESS	catherine.mcclellan@acer.org
DIRECT TELEPHONE	+61 3 9277 5249
MOBILE	+61 (0) 438 154 194

CONTENTS

1. Executive Summary

This report considers several aspects of the National Assessment Program – Literacy and Numeracy (NAPLAN) writing assessment, with a specific focus on the marking. The first component is a qualitative review of the marking rubric itself, as well as aspects of the design of the writing assessment, consideration of alternative approaches, with limitations and issues to be considered before making changes. The second component is primarily quantitative analyses, including factor analytic and item response theory approaches, with an emphasis on examining the prevalence of local dependence in the data. The third component of this review comprises an examination of the marking design and includes a review of current and potential marking quality assurance measures. The final section outlines a program of research and study to support progress in program improvement.

This work was done in consultation and collaboration with a range of people with expertise in relevant aspects of the review, such as content developers, assessment designers, psychometricians, policy professionals, marking specialists, and those with deep practical knowledge of existing operational procedures. This multi-faceted view was sought with the intent of providing useful and realistic guidance on the continuing improvement of an already-strong NAPLAN writing assessment program.

Key conclusions of the text include:

- The NAPLAN writing rubric is unusual in that it comprises 10 criteria with the intent of providing useful formative information to teachers, students, parents, and other stakeholders.
- There are specific features of the criteria and overall rubric structure that may benefit from reconsideration, including the way the score categories are structured, the relative weighting of different aspects of writing skill, the complexity and language of the descriptors, and overlap in the use of evidence across criteria.
- Aspects of the assessment design interact with the rubric and criterion considerations. Sampling a single essay in one text type limits the generalisability of the assessment results, and does not span the full range of text types in the Australian Curriculum: English. An alternative model is discussed, with some advantages and limitations noted, and approaches to assessing writing in other international contexts are briefly outlined to provide some context.
- The quantitative analyses pointed clearly to two major findings. First, the frequency and use of score categories of the criteria is inconsistent, with infrequent categories occurring in nearly every criterion and year level. And second, every analytic approach consistently and strongly pointed to significant local dependence issues in the data that were not remediated by any analysis completed herein.
- The evaluation of the marking design resulted in recommendations for marker training, exemplar selection and classification, and quality assurance measures. The current operational approaches to these are outlined and, where appropriate, modifications suggested for consideration. The implementation of automated essay scoring systems is briefly discussed, with a particular focus on reducing burden on human markers and focusing their work on the aspects requiring the most professional judgement.
- The last section contains brief descriptions of a series of research studies targeted at developing the evidential bases needed to support planned changes and improvements, with consideration of dependencies and risks for each. The results of the studies would guide

progress towards an assessment program that has a clear and limited purpose; one that provides useful and actionable results that can inform and guide instruction and improvement; one where the intended uses for scores have clear and compelling validity evidence cases; and one where the designs for the assessment, the marking rubric, and the work of operational marking are carefully and thoughtfully planned to work together seamlessly.

This report has been informed by a broad and varied set of viewpoints, but it does not encompass every possible appraisal of the NAPLAN writing program. Stakeholder opinions have been considered where available, but the predominant views come from experts who work inside the assessment field. As a result, the report optimises suggestions and recommendations that address concerns most prevalent from those perspectives. Use of constructed-response assessment such as writing is an ever-evolving field, and this fact combined with the program of research recommended herein could results in new advice that supersedes the current. Just as with the NAPLAN program itself, this report is a snapshot of the current moment, and will always be subject to improvement as knowledge increases and methods improve.

2. Qualitative Rubric Review components

2.1. Category Labels and Descriptors

The first aspect of the qualitative part of the rubric review is the category descriptors. These are both an obvious target and relatively low cost to alter, making them the “low-hanging fruit” of rubric reviews. NAPLAN’s rubrics have been revised and structural changes made over time, a factor that is not uncommon in long-lived assessment programs. The changes were carefully thought-out and planned. The cumulative impact of these changes may have resulted in some unintended effects.

Marking rubrics are often the end result of a committee and consensus process. Even if they are initially created by the same experts who develop the prompt materials, they may be constructed with a longer view in mind for programs like NAPLAN. After all, the rubrics are one component—sometimes the only one—that remains stable in an assessment system. Specific topics change every year. The writing text type is selected annually, but is not the same for more than a few years as a rule. Markers come and go, with new ones in varying proportions annually at each centre. Marking team leadership changes as experienced markers retire and new ones advance to the roles. Scoring notes, formal and informal, appear during every live marking window as student submissions change and may or may not be retained across administration cycles. Even marker training varies somewhat with delivery by each individual trainer. But the rubric (mostly) lives on through all those alterations, as perhaps the most stable facet of the assessment, and thus contributes in important ways to the test’s reliability and validity of inference made from the results.

2.1.1. Weighting

The most distinctive element of the NAPLAN rubric is its relatively extensive set of criteria. As the table at the end of this section shows[1], the use of 10 criteria for a high-stakes writing assessment is unusual in international education. It appears there is little evidence that other countries use as many criteria; the closest is New Zealand with seven criteria, but the assessment is computer-based and low stakes. It is accepted that the aim of having a 10-criteria analytical marking scheme was to test the breadth of skills needed in effective written communication. Such rubrics ensure that the markers take each criterion into account and theoretically might encourage greater marking consistency. However, consistency is contingent on having ‘air-tight’ descriptors that are understood in the same way by all markers and applied by each marker as intended.

An examination of the 10 criteria used in the NAPLAN Writing assessment shows that the breadth of writing skills is comprehensive, but the weighting of the skills might be reconsidered. Assessments of writing which view writing as a set of discrete skills usually focus on the more technical elements of writing, such as conventions of grammar, punctuation, structure and spelling. Assessments of writing that view the target as a more integrated construct are likely to include aspects of compositional or authorial skills in the set of assessment criteria, such as ideas and the ability to write for a particular purpose or audience. It appears that the NAPLAN criteria were constructed to take both of these aspects of the writing construct into account, but in quantifying the skills into a scoring scheme, the technical skills outweigh the compositional qualities.

There are a number of ways to partition the NAPLAN rubric criteria into groups. As one example, in the NAPLAN narrative rubric, out of a potential total of 47 score points, 15 of the points are allocated to the criteria reflecting the traditionally-defined compositional aspects of writing (audience, ideas,

[1] Ofqual (2019), Coventry Ofqual report: A review of approaches to assessing writing at the end of primary education

character and setting). If the remaining criteria, representing aspects of writing such as organisation and conventions, are combined, there are 32 score points awarded based on these more-technical aspects of writing.

Some writing experts would argue for the inverse weighting of these criteria, as the compositional aspects allow for determining a more significant discrimination in the writing ability of students. Good writing is always more than the sum of its technical parts and it is possible to reward this quality of writing through a greater weighting of the compositional aspects. This can be done by re-structuring or revising the rubric itself, or it can be done analytically post hoc. Either approach would require careful consideration of both the desired weighting to be applied and the practical and policy implications of this re-structuring.

2.1.2. Rubric Design and Use

If the rubric for a criteria relating to the content of writing is carefully constructed and has detailed score categories, then markers are able to score this aspect of writing consistently. An example of a detailed rubric for the content of a narrative piece of writing is seen below in Figure 1[2], where there are 10 score categories in total (note: only the top five score categories are shown here), with detailed descriptors for each category.

LEVEL	DESCRIPTION	POINTERS
10	The **writing is sustained and presents a complex and mature reflective viewpoint or approach**. If a narrative, the writer may adopt and sustain a convincing persona as author or participant in the action. The skilfully constructed piece displays originality and is supported by carefully selected detail. If a narrative, characterisation shows emotional or psychological complexity. The writing evokes a strong response in the reader.	• sustained narrative / reflection with complexity of purpose, sophisticated approach or subject matter • thought-provoking reflection on attitudes, values or issues • writer adopts convincing persona • emotional or psychological depth to characters • evokes strong response in reader • likely to go beyond stereotypes
09	**A carefully constructed piece that may reflect on values and offer insights**. There may be some reflection underpinning or implicit in the piece. If a narrative, characters are credible, with the reader given insight into their lives. Relationships between characters are convincing.	• sustained and unified /reflection with a well-constructed conclusion • may be more than a linear construction, for example, more than one complication • empathetic response to characters • reflects on attitudes and values • reader's interest strongly caught
08	The writing **is a developed piece**. The **overall structure is appropriate with a clear direction.** (may be unfinished) If a narrative, characterisation is credible, for example, through the presentation of motive underpinning action or emotional response to the situation. If a reflection, the writing conveys genuine engagement with the task and the reader.	• developed and integrated • credible character development/reflection • sure as a narrator • attention to time order • engages reader
07	The writing **is a well-constructed piece within a sound structure.** A deliberate intention of engaging the audience is evident. Ideas and events are appropriately linked. If a narrative, the characterisation is credible, with characters clearly individualised. If a reflection, the writer's point of view is clear.	• sound structure • sense of voice • generally sound characterisation (narrative) • clear point of view (reflection)
06	The narrative / reflection has a **clear sense of purpose** The writing contains ideas, details and events chosen to enhance the piece of writing. Characters are distinguished either explicitly through description or implicitly through action and speech	• focus maintained • characters clearly defined • a degree of detail • sense of audience

Figure 1: Sample writing rubric (part thereof) for assessment of content

The detail in the content rubric in Figure 1 illustrate a case where audience, narrative devices, cohesion, and ideas (all current authorial criteria in NAPLAN writing) have been collapsed into one 'content' criteria.

Whether rubrics are complex or relatively simple, it is clear that when making judgements about student writing, markers vary in their adherence to marking rubrics and/or assessment criteria, and can make relative, as opposed to absolute, evaluations of students' work. Effectively this becomes a 'localised' version of pairwise comparisons within each marker. While pairwise comparisons have been shown to be very effective in constructing a reliable classification scheme for writing[3] at the

[2] Rubric is from the International Schools Assessment (ISA); other criteria used in the assessment include language and structure, not shown here. Rubrics are included in Appendix A in their entirety.

[3] See, for example, Humphry & Heldsinger (2014, 2013) or Humphry & McGrane (2014).

holistic and criterion level, this internal and informal approach lacks the structure that will assure sufficient judgements are applied to each essay.

2.1.2.1. Zero as a Score Level

One obvious problem with having so many criteria is the potential lack of uniformity of interpretation by markers of the category labels and descriptors. Currently, the category zero (0), which exists for each NAPLAN criterion, is problematic because there is inconsistency in the descriptors. For example, the descriptor for category 0 in the Audience criterion for Persuasive writing states: *symbols or drawings which have the intention of conveying meaning.* However, the category 0 description in other criteria for persuasive writing states: *no evidence or insufficient evidence*. For one criteria for a score of 0 there is 'something' that is assessed (symbols and drawings) and for another criterion, there is 'nothing' assessed. Such inconsistency in the descriptors of scores undermines the scoring process.

In terms of measurement, it is recommended that score levels of 0 are avoided unless there is 'nothing' on the page to score (i.e. *no evidence*). Human markers have an aversion to assigning a score level of 0 to writing with any discernible content, as the connotation of 'nothing' is quite strong. It is preferable that responses with no content be categorised into 'blank' for responses with no text and assigned a special code to indicate the missing data. For responses with 'no discernible text or comprehensible words' such as symbols, drawings, complete erasures, or random letters, the zero score level could be retained.

If it is retained, it should be described in the same way across all criteria. If there is no discernible text for Audience, there is no discernible text for Spelling or Sentence Structure or any of the other criteria as well, so assigning 0 to any criterion should imply a 0 for all criteria. In this way, the interpretation of a mark of 0 is consistent across all criteria. It is not an arbitrary judgement about how little there is of something. Consistency of interpretation in the meaning of the score level will be held stable across the full set of criteria, and thus also in the use of the category by the cohort of markers. It is suggested the scoring of actual writing content would start at category 1, and this score would describe the lowest level of achievement for the given criteria. This is a recommended change to the marking system. Note that implementing this approach will introduce a discontinuity in the score distribution: the next-lowest score above 0 will be 10 and scores 1-9 will not be possible, assuming the current number of criteria are retained.

A further example from the Senior Marker Compendium[4] highlights the issue of (in)consistency. Here are two reports on the award of a score of zero for spelling for two quite different pieces of work:

1.

10. Spelling	**0**	Some simple words may be distinguishable (*he, to, the, all*); however, because text is predominantly letter strings, there is a lack of context to verify meaning.

2.

10. Spelling	**0**	No evidence.

In the former, which contains quite a bit of writing and some correct words, the score is 0 for Spelling. In the second piece, where the writing has all been erased, the score is also 0.

[4] ACARA, 2018; page 2 and page 4

2.1.2.2. Quantification

Another potential problem with the current NAPLAN rubric is that, for certain categories, it encourages a specific count of the skill being assessed. This quantification of the skill may lead markers to think the number of times the skill is displayed is more important than the substantive evidence of the actual competency itself. For example, in the Text Structure criterion for Persuasive writing, a specific number is nominated for a score of 2: the *text contains **two** clearly identifiable structural components.* Likewise, in the Persuasive Devices criterion, category 2 states: *uses **three or more** instances of persuasive devices.* In this latter example, it is well-understood that some persuasive devices are more sophisticated, or more subtle, than others. Yet, if a student does not have the requisite number of devices according to the descriptor in the category (but has used one or two sophisticated devices with aplomb), does this mean that student cannot receive a score that reflects their level of competency in using these persuasive devices? These are just two examples of the 'counting' phenomenon that exists in some of the current descriptors, which may be working against the stated aim of a reliable marking rubric that produces scores supporting valid inferences about writing ability.

Another factor that should be taken into account by stakeholders in relation to the descriptors in the categories is the effect that this counting and quantification can have on the time taken by markers to score these categories. Whenever a 'count' is required by markers, this extends the time taken to mark the piece of writing. This has implications for the costs incurred and for the marking timeline itself, as a longer marking window is needed. While it may seem that counting specific instances would be less time-consuming than the more sophisticated task of judging quality, that is not always so. Despite the apparent clear-cut nature of many of the counting criteria, some are more subtle, as the Text Structure descriptions above indicate. Especially where interpretation is required in order to decide if a particular piece of text qualifies within the meaning of the score level to be counted, it may result in more conferring and discussion. Markers may become uncertain and think, 'I'll just check with my lead marker to see whether I should count this as a complex sentence or not'. Counting often interrupts the flow of marking quite noticeably and this stop-and-start pattern increases the marking timelines.

Experience shows that 'counts' may be included in rubrics to circumvent the problem of vague qualifiers in the descriptors that are intended to assist markers in making a decision about a score. There is nothing inherently wrong with using qualifying words in descriptors and they are commonly relied upon in rubrics to differentiate between score categories. However, currently in the NAPLAN rubric, there appear to be inconsistencies in the way they are used. For instance, in the Ideas criterion for Persuasive writing, the quantifier 'many…ideas' appears in score categories 2 and 3, but does not appear in score categories 4 and 5. This is potentially confusing for markers. For a score of 3, the rubric states: *many unelaborated ideas that relate plausibly to argument (four or more).* Yet for a score of 4 in this category, the rubric simply states: *ideas are elaborated and contribute effectively to the writer's position*. There is no mention of the qualifier 'many' at the higher score levels, so taken at its most literal, the lower-scored responses need 'many' ideas, while the higher-scored responses do not. This is one example of contradictory, or at least confusing, wording of descriptors.

Markers may be trained to treat the criteria as cumulative in the sense that the category descriptors build on one another. If this is so, a quantity specified at a lower score level would be implicitly required as a minimum at all higher levels. If this is the expectation for how the rubric is used, it would be better to state this explicitly in the criteria, removing reliance on each trainer to cover this content and on markers to recall it while marking.

Other vague qualifiers that are used in the rubric include ‘few’, ‘some’, and ‘range’. As stated earlier, there is nothing inherently wrong with using such qualifiers, but what must be clear is the markers understanding of the gradations that each of these words implies. It is important that there be consistency of the qualifiers across criteria, not only within a criterion. The terms should mean the same thing for all categories for all criteria. Emphasis on this in marker training and provision of examples that explicate the meaning of the quantifier is thus very important if the descriptors are to be interpreted consistently.

2.1.2.3. Parallelism

NAPLAN’s writing rubric places high cognitive demand on markers. Across 10 criteria, there are 57 (Narrative) or 58 (Persuasive) score categories that markers must understand and hold in short-term memory roughly simultaneously as they read the text. This total counts the zero level as a score category, since as noted above it is not consistently defined across criteria. Even if 0 is eliminated, there are 47 or 48 categories. The cognitive load imposed by this is not trivial. An aspect of the NAPLAN rubric that could be reviewed is the consistent use of the same phrases or descriptors across the score levels as well as the different criteria to reduce the memory and processing load demanded for use.

One way the cognitive load is increased is through the use of different terminology and descriptors in some criteria across score levels. For example, for Cohesion on the Persuasive rubric, the level 1 descriptor includes *links are missing or incorrect*. Level 2 states *some correct links between sentences*, at level 3, *controlled use of cohesive devices* and at level 4, *range of cohesive devices is used correctly and deliberately*. Between level 2 and level 3, links shifted into cohesive devices. It is unclear if these are meant to refer to a similar or the same underlying skill or how they are connected. It is also tautological to use the word ‘cohesive’ in the definition of the criterion ‘Cohesion’.

Introduction or removal of skills or concepts partway through a criterion also can generate additional cognitive demand. For example, in the Audience criterion on the Narrative rubric, the level 2 descriptor is *shows awareness of basic audience expectations through the use of simple narrative markers*. The phrase ‘audience expectations’ is absent from the levels 3, 4, and 5 descriptors, but a version of it appears again at level 6 (*caters to the anticipated values and expectations of the reader*). In level 6, it is augmented with an expectation around anticipated reader ‘values’—a term appearing in the criterion for the first time at score level 6. In Text Structure, a lack of time-sequencing is called out as a reason for the assignment of a score of 0, implying it is problematic, but at no other level is time-sequencing indicated as a desirable facet of the text or reason to assign a score level.

Making score level descriptors parallel, in that they refer to one set of underlying skills using the same terminology, allows markers a simpler path into understanding and using the rubric effectively.

2.2. Evidence Structures

Utilisation of the same evidence to align a response to a score level in more than one criterion can undermine the effectiveness of the rubric through an increase in dependence in the scored data (see Section 3 for more details about data dependence). It can be confusing for markers if criteria are so similar that the same evidence can be applied, and it leads to a system that rewards or punishes a student repeatedly for the same action. To paraphrase a Russian idiom, as currently configured the rubric forces students to keep stepping on the same rake. It is preferable that evidence be aligned appropriately in only one criterion, or at least that there is a clear single best choice of criterion for where specific evidence of writing skill belongs.

An example of how this can be perplexing for markers can be seen where the rubric states, for the

criterion of Cohesion (score 1): '*short script*. In the criterion of Vocabulary (score 1) it also states: (*very*) *short script.* Confusingly, the descriptor *text is very short* is used as 'Additional information' in the criterion of Audience, for a score of 1. If the categories are meant to be separate, and Vocabulary, Cohesion and Audience are three distinct aspects of the writing construct, then it follows that the same descriptor/phrase should not be used in more than one criterion. Here, the length of the script is deterministic for scoring 3 (nominally) different criteria; this appears to defeat the purpose of having separate criteria.

Consideration of the same evidence as 'counting' for more than one criterion on the rubric leads to violations of local independence that will cause statistical problems in scaling the assessment results. It also can heighten overly-frequent use of the same score patterns often observed in scoring of complex performances.

2.2.1. Potential Redundancy

The volume of technical criteria contained in the writing rubric should be considered against the skills already being tested in the NAPLAN suite of assessments, in particular in the Conventions of Language (CoL) test. Spelling, punctuation and grammar (including sentence structure) are all tested in Conventions of Language, so there is an apparent over-testing of these particular abilities. It is important to note that in the NAPLAN Writing assessment, students are presented with a requirement to produce, edit, and correct their own original text, in contrast to the CoL assessments tasks, which are more editorial and generally require interacting with provided text. This distinction may mean that the skills demonstrated on the two assessments are not the same. If the language skills utilised in these different parts of NAPLAN are shown to produce equivalent results, students are being repeatedly penalised or rewarded for what will be the same outcomes on a reliable assessment. If these skills on CoL and Writing are not meaningfully distinct, students are being burdened for what may be superfluous information about their status.

In the NAPLAN assessment framework (NAPLAN Online 2017-2018), it states 'the NAPLAN writing test aligns with the Australian Curriculum: English through a focus on the following sub-strand threads: Purpose, audience and structures of different types of texts, Vocabulary, Text cohesion, Sentences and clause level grammar, Word level grammar, Punctuation (and) Spelling'. Likewise, the framework describes the Conventions of Language test as focusing on 'the accurate knowledge and use of the spelling, grammar and punctuation conventions of Standard Australian English'. It seems that the framework description for NAPLAN Writing focuses on the same technical aspects of writing and covers the same ground as the Conventions of Language test.[5] Further clarification of whether there are measurable differences in the outcomes when students use their spelling, grammar, and punctuation skills to produce original text, in contrast to correcting or completing provided sentences, would be useful in evaluating the extent of possible redundancy in NAPLAN assessments.

The range of issues described in Sections 2.1 and 2.2 suggest that a thorough and detailed audit of the NAPLAN rubric would be productive in improving clarity and simplicity of use for the markers. All descriptors and criteria should be reviewed with these issues in mind, and the samples and compendia carefully aligned once changes have been agreed.

2.3. Single-Task Design of Writing Assessment

[5] The Australian National Assessment Program Literacy and Numeracy (NAPLAN) assessment framework: NAPLAN Online 2017-2018, pgs. 12-14

One of NAPLAN's major goals, as well as an uncommon feature for a large-scale, standardised assessment, is the provision of fairly detailed descriptive information[6] to teachers and students. This intention is confounded with a single writing sample design. Students can be assessed in only one text type of the two currently active in NAPLAN, and there are text types that are not assessed even though they appear in the Australian curriculum. As noted on the NAPLAN writing website:

> *The Australian Curriculum: English requires students to be taught a variety of forms of writing at school. The three main text types (previously called genres) that are taught are imaginative writing (including narrative writing), informative writing and persuasive writing. In the writing tests, students are provided with a 'writing stimulus' (sometimes called a prompt – an idea or topic) and asked to write a response in a particular text type. Students are tested on either narrative writing or persuasive writing. Informative writing is not yet tested by NAPLAN.*[7]

This places limits on the generalisability of the assessment. Addressing this issue would require extensive changes to the current NAPLAN writing assessment design that would have to be considered and implemented carefully over an extended period of time.

There are advantages and disadvantages of a single writing task with multiple criteria for assessment, including the time required, potential dependence when marking a single essay across criteria, and the range of text types that can be included.

2.3.1. Testing time

An obvious advantage of a single writing task is that it can be completed in a relatively short time: with 40 minutes, students can reasonably be expected to plan their writing to some extent, produce an assessable artefact, and possibly give a light proof-read. An obvious disadvantage of long writing tasks, conversely, is the very fact that they take time, which is always a critical factor in large-scale assessments, exacerbated in mandatory population tests, and further exacerbated in the context of testing young children.

Particularly for the youngest children in the NAPLAN context, eight- and nine-year-olds, there is a question about whether 40 minutes spent on a single task is the most effective way of gaining maximum information about a child's level of achievement in an essential area of learning. Aside from anything else, if the topic does not appeal to the student, there is nothing else to compensate. There is a well-known aversion to 1-item tests in the assessment field, for good reasons[8]. Multiple criteria on a rubric do not fully compensate for only a single essay marked against them.

2.3.2. Dependence among criterion scores

Even where there are several rating criteria, each yielding several score points, since only one artefact is available for assessing writing, there is inevitably a high degree of dependency among the scores. Setting aside the measurement evidence that indicates a high degree of dependency among the criterion scores (explicated further in Section 2), from a qualitative perspective this is predictable, and the more so the greater the number of criteria. NAPLAN's establishment of 10 criteria on which to assess a single piece of writing is extremely unusual, if not unique. Experience of marking itself points

---

[6] 'The tests provide parents and schools with an understanding of how individual students are performing at the time of the tests. They also provide schools, states and territories with information about how education programs are working and which areas need to be prioritised for improvement.' (NAPLAN frequently asked questions: general).

[7] NAPLAN writing website

[8] Studies of the effects of test length go back to Spearman (1910) and are included in books such as Gulliksen (1950) and Lord & Novick (1968) and articles including Bell & Lumsden (1980) and Fitzpatrick & Yen (2010) among many others. An accessible discussion is provided in Livingston (2018).

to the difficulty of distinguishing what is uniquely attributable to individual criteria: for example, disentangling the relationship between 'ideas' and 'persuasive devices', or 'text structure' and 'cohesion'. Attempts to ascribe distinct scores to each criterion are likely to lead to factitious directions to markers, such as counting the number of ideas, regardless of their quality.

The implications for teaching and learning are not desirable. The current design of the NAPLAN writing assessment seemingly provides rich formative information, with marks for 10 different aspects of the piece of writing. However, teachers, students, and carers may not understand the strength of the relationships between these aspects, and may perceive them as more distinct and specifically diagnostic than the data can support. The value of each criterion score probably is more like a set of views of a unitary proficiency in writing, each taken from a slightly different angle. The criteria scores should not be viewed as neatly partitioned elements delineating the strengths and weaknesses of the individual student's performance.

2.3.3. The range of text types assessed

The Australian Curriculum refers to 'imaginative, informative and persuasive' text types[9] but these are clearly very broad categories, each comprising a wide range of sub-types. The NAPLAN writing assessment has operationalised only two of the three broad categories (imaginative and persuasive), and within each of these a similar style of prompt has been used, with a single formulaic set of guidelines accompanying the prompt. It should be acknowledged that these are 'guidelines', not directions, and that the predictability of the prompt styles and the guidelines are undoubtedly reassuring for both teachers and students; they also unquestionably provide support and scaffolding, especially for struggling students. These are important and positive aspects of the current approach. However, the uniformity of the writing tasks over the years also can have a stultifying effect on the learning and teaching of writing. 'Teaching to the test' is not a bad thing when the test represents well the construct we want taught, but when the test is evidently narrow in its approach, the consequences are likely to be damaging. Echoes of this view are apparent in the Perelman review[10] of the NAPLAN writing assessment. Wyatt-Smith & Jackson (2019) report that educators indicated the NAPLAN writing assessment directed policy and practice focus onto writing and generated opportunities for professional development in supporting improvement in student writing, with the marking criteria viewed as an indicator of specific areas of concern to be targeted in instruction. Study participants indicated that writing in a variety of genres was an area of opportunity for NAPLAN.

2.4. An alternative model

A different model of assessing writing, which overcomes some of these issues, comprises several short writing tasks. Section 2.6 includes a brief description of several international assessments of writing, including some that incorporate multiple writing tasks and genres. Each task is rated on a relatively small set of criteria. Across a set of, say, three 10- to 15-minute tasks of this kind, with different text types, each yielding several score points across a limited number of criteria, a scale of 20 to 35 points can be generated. The advantages of this kind of assessment are:

- Assessment of a range of text types;
- Yield of multiple, relatively independent score points from a writing assessment of 40 minutes; and

[9] Australian curriculum: English, key ideas
[10] See ABC news story: NAPLAN's writing test is 'bizarre' but here's how kids can get top marks for a summary.

- Coverage of different and distinct writing criteria, matched to the task type.

Examples from a writing assessment for upper primary with this kind of design, the Monitoring Trends in Educational Growth (MTEG), are provided below in Figure 2[11].

Writing Task 1

The Bird and the Box

Write a story about this picture.

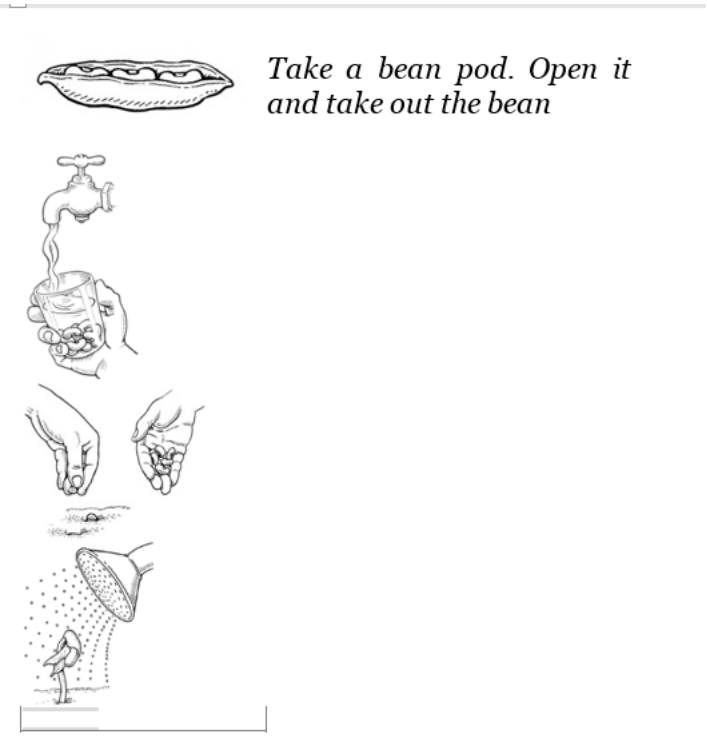
Write as much as you can, on the lines below. Try to make your story interesting. (10 lines provided for writing.

[11] The assessment shown is from Class 6 Writing Assessment, MTEG Afghanistan (ACER and Afghan Ministry of Education, 2013). The Writing Assessment in the example was one of six forms rotated across the sample group. The total writing task pool comprised 11 tasks of varying length and difficulty. See MTEG website for more details.

Writing Task 2

How to Grow Beans

Write instructions for planting and growing beans. Us the pictures to help you. The first one has been done for you.



Writing Task 3

Visiting Cousin

After school you come home but no-one is in your house. You decide to go to your cousin's house nearby. Write a note for your brother to explain what you are doing. Write three or four sentences.

Dear brother, [Seven lines provided for writing]

Figure 2: Example of a multi-task writing assessment

The writing assessment shown in Figure **Error! Reference source not found.**2 comprises three t asks which are to be completed in 30 minutes. Within the current NAPLAN time allowance of 40 minutes for the writing assessment, slightly more time could be allowed per task.

Table 1 below shows the text type, marking scheme and time allowance for each of the three tasks.

Table 1: Features of a multi-task writing assessment

Writing task	**The bird and the box**	**How to grow beans**	**Visiting cousin**
Estimated time per task (minutes)	15	7.5	7.5
Text type	Narrative	Instructional	Transactional
Marking criterion 1 (maximum score)	Narrative sequence (2)	Instructional language (2)	Ideas / relevance (2)
Marking criterion 2 (maximum score)	Elaboration of ideas (4)	Relevant information (2)	Vocabulary (2)
Marking criterion 3 (maximum score)	Punctuation (2)	Spelling (2)	Handwriting (2)
Marking criterion 4 (maximum score)	Sentence structure and complexity (3)	NA	NA
Total maximum score	**11**	**6**	**6**

The content of the tasks and the nature of the assessment criteria are clearly different from those that would appear in an Australian assessment of writing (for example, in Writing task 3, the reference to 'Dear brother', and the criterion 'handwriting'). Nevertheless, the assessment has several features that could be incorporated into NAPLAN writing.

1. *A range of text types*: Three text types (narrative, instructional and transactional) are presented in the example. The full task pool for the MTEG Year 6 assessment descriptive, persuasive, informational and labelling tasks, with length of text required ranging from single word, sentence level, lists and more extended (half- to one-page pieces). A design including potential selection from wide range of text types would promote a wider and more flexible approach to the learning and teaching of writing, as well as providing students with the opportunity to demonstrate the scope of their writing capacity.
2. *A range of criteria to provide formative information*: The criteria cover ideational content, micro and macro structure, and linguistic features. The marking scheme as a whole aims to give significant weight to each of these three elements across the combined tasks – aiming for a balance of content, structure and technical appropriateness.
3. *Avoidance of dependency among criteria*: In addition to reducing dependency by the inclusion of multiple tasks, the number of criteria assessed per task is also a way of minimising dependency. There is a small number of criteria for each task, assigned with the aim, not just of reducing the number of judgments per task, but also of sufficient differentiation of skills to ensure that each criterion is judged independently within the individual task.

2.4.1. Equating across year levels

A writing assessment design similar to that described above has been implemented by ACER in a number of programs, including the New Zealand Literacy and Numeracy for Adults Assessment Tool, and the writing assessment for Year 5 students in the regional South-East Asian Primary Learning Metric, as well as the Monitoring Trends in Educational Growth assessment for Year 6 students from which the example above was drawn. All of these assessments focus on a single year group (*e.g.* Year 6) or cohort (*e.g.* Adults). A further advantage of a multi-task model of writing assessment for a program across different year levels, like NAPLAN, is that linking via tasks (comprising prompt and marking rubric) can be obtained by including common tasks across year levels, on a similar model to that used for reading and numeracy assessments. To illustrate, building on the example from Figure 2, the design shown in Table 2 could be used across the four year levels of NAPLAN writing:

Table 2: Model of a multi-task writing assessment across NAPLAN year levels

Year 3	**Year 5 (as shown in Figure 2)**	**Year 7**	**Year 9**
Basic task (*e.g.* labelling)			
Less challenging task than *The bird and the box* (*e.g.* descriptive / imaginative – based on a picture)			
The bird and the box (narrative)	**The bird and the box (narrative)**		
How to grow beans (instructional)	**How to grow beans (instructional)**	**How to grow beans (instructional)**	
	Visiting cousin (transactional)	**Visiting cousin (transactional)**	
		More challenging task (*e.g.* narrative)	More challenging task (*e.g.* narrative)
			Most challenging task (*e.g.* persuasive)

A design of this kind obviates the need to rely solely on the marking rubric as the tool for vertical equating. If marking is designed so that markers score essays in year-level-specific batches, it is doubtful that they are able to maintain a constant frame of reference across the year levels. NAPLAN rubrics assume that a score of 1 on a criterion for Year 3 means exactly the same thing as a score of

1 on the same criterion for Year 9 and this underlying assumption is enacted in the current long scale equating for writing.

2.5. Issues to be considered if a change to the writing assessment design is implemented

2.5.1. The 'demise' of sustained writing

A disadvantage of adopting a multi-task design for the assessment of writing that fits into the current 40 minute time allowance is that it may be perceived as discouraging the kind of sustained 'thinking in writing' that the demand for an essay-length written response promotes. Given the wash back effect of NAPLAN on the curriculum as a whole this is a reasonable concern.

The objection to a design that includes longer pieces of sustained writing could be mitigated by including just two tasks at Year 9 (as shown in Table 2), each allowed 20 minutes. This would still allow for greater variation of text type, and a smaller set of criteria for each task with less dependency, giving continuing formative information to students and teachers. There is an established body of research on the impact of reducing the time allocated to extended writing tasks. Broadly speaking, reduction in time led to essays that were shorter and with some impact on quality, but the essays tend to be equally complex to those written in a longer time window. Rank ordering of essays tends to be stable in the face of time reduction as well. A graduated increase in the length and concomitant complexity of the writing task would pave the way for senior secondary writing expectations. An alternative option worth considering would be to make the Year 9 writing time allocation greater so that there could be multiple tasks with more time allocated to each.

It might also be noted in this context that the NAPLAN reading assessment – like all large-scale reading assessments –is based on short pieces of stimulus: often a mix of short self-contained texts, such as poems or instructions, and extracts from longer texts such as novels. NAPLAN's test design for reading has not led to the death of reading of extended works such as story picture books and junior fiction in primary schools, or complete novels and plays in lower secondary. While NAPLAN's influence on Australian teaching is undoubtedly broad and deep, it does not fully constrain classroom instruction in reading and it is unlikely it will do so in writing if these changes were implemented. The advantages to be gained from assessing writing via a more diverse range of text types, in our view, outweigh the single-prompt design.

2.5.2. Concomitants of any radical change

As with all changes in a very public domain the change proposed would undoubtedly cause controversy because is it different from what the educational stakeholders have become accustomed to. It will also create challenges for measurement *a la* the classic quote: 'if you want to measure change, don't change the measure.' As stated earlier, addressing the issues of text type and generalisability would need to be implemented carefully over an extended period of time. One approach for doing so will be outlined in Section 5.

2.6. Addendum: Other Writing Assessments

Writing in most Australian standardised assessments has typically and traditionally taken the form of one or two extended writing tasks. This the case with, for example, the current version of the GAT[12], the Western Australian OLNA, the New South Wales Minimum Standard Writing Test, the ACT's Australian Scaling Test and the Graduate Medical School Admissions Test (GAMSAT), as well as NAPLAN. Some of these assessments are marked to produce a single holistic score per task (for example, GAT and GAMSAT). From a measurement perspective, several scores will give a more reliable result than one score; for high-stakes tests where only one or two writing tasks are marked holistically, the reliability of a single mark per task is improved by having several markers rate each piece. In the case of GAMSAT, for example, three markers read each script.

The Ofqual publication *A review of approaches to assessing writing at the end of primary education*[13] provides a useful survey of English-language national and sub-national large-scale assessments of writing. Of the 15 assessments described, several are similar to the current form of NAPLAN's writing design, based on a single extended response (Ontario, Hong Kong, New Zealand), or two extended response tasks (USA – National Assessment of Educational Progress; Singapore). Others rely on portfolios of writing assessed by teachers with moderation (England, Caribbean). A third group assesses writing only through multiple-choice items (Scotland, Philippines). Two Californian assessments, one online and one paper-based, use a mixed model, multi-task approach.

[12] Note that, at the time of writing, the new GAT writing assessment, currently under development, is planned to comprise a 30-minute argumentative task, and two or three short tasks to be completed in 30 minutes.

[13] Table adapted from Ofqual (2019), Coventry Ofqual report: A review of approaches to assessing writing at the end of primary education

Jurisdiction	Assessment	Method of the writing assessment
Australia	NAPLAN	Extended response type items. Traditionally paper-based, but a sample of students were tested online in 2018; marked on 10 criteria: Audience, Text Structure, Ideas, Persuasive Devices/Character and Setting, Vocabulary, Cohesion, Paragraphing, Sentence Structure, Punctuation, Spelling
Canada (Ontario)	Junior Division Assessment; JDA	Paper-based, with extended response and multiple-choice type items; marked on two criteria: 'topic development' (6 levels) and 'conventions' (5 levels)
Singapore	PSLE	Two pieces of writing in 70 minutes; paper-based. 1 piece of 'situational writing' which constitutes a short 'functional piece' (letter, email, or report), and 1 piece of 'continuous writing' which constitutes a longer (150 words minimum) piece of continuous prose based upon a given prompt. Students are marked according to 2 domains: 'content' and 'language and organisation'.
New Zealand	e-asTTle	Computer-based test; 20 prompts are available, covering 5 writing purposes (describe, explain, recount, narrate, persuade), from which teachers choose 1 piece of extended writing in response to this prompt, with a time limit of 40 minutes, Marked separately on 7 domains: ideas, structure and language, organisation, vocabulary, sentence structure, punctuation, and spelling.
United States of America (California)	English Language Proficiency Assessments for California (ELPAC)	Paper-based, with a mixture of item types: short responses (1 or 2 sentences) and longer extended responses (1 or more paragraphs); focussed criteria on describing a picture, writing about academic information or experiences, or justifying opinions.
United States of America (California)	California Assessment of Student Performance and Progress (CAASPP)	Computer-based, with a mixture of multiple-choice, alternative format (*e.g.* clicking on sections of text), single paragraph and multiple paragraph extended-response items; marking criteria focus mainly on writing for a purpose (*e.g.* developing narrative, presenting evidence), with a limited focus on technical writing skills

United States of America: NAEP	Computer-based test	Each pupil completes two 30-minute extended-response type tasks in response to a given prompt. In each task, the intended audience of the writing is clearly stated/implied. Levels-based holistic marking scheme, supported by level descriptors, is used to give each pupil a single score of 1-6 for each task.
Scotland	SNSA	Computer-based test; Writing questions target spelling, grammar, and punctuation only; assessments are marked automatically online.
Trinidad & Tobago	SEA	Paper-based test; test contains either 3 narrative (story) items, or 3 expository (explanatory) items; externally double-marked (holistic) on content, language use, grammar and mechanics, and organisation.
Hong Kong	TSA	Paper-based; an extended piece of writing of about 80 words (*e.g.* a story or a letter), based on a given prompt, in about 25 minutes; marked out of 4 for each domain: content (level of detail and clarity) and language (*e.g.* vocabulary, verb forms, grammar)
England	KS2	Portfolio; internally assessed by teachers, a sample of which are externally moderated.

Table 3: Summary of international writing assessments

3. Quantitative Rubric Review components

3.1. Data Summary

The data used for this study are 2018 and 2019 NAPLAN Writing data. For all analyses described in this section, the Stage 2 census data was used (i.e. the version of the data that was used for the national report). Two test mode were used in these two years, pen and paper test and computer delivered online tests. The majority of students in 2018 did the NAPLAN paper tests, and only a small proportion of students in 2018 did the NAPLAN online tests. In 2019, about 50% students did the NAPLAN tests in paper and others did the NAPLAN online tests. Table 4 shows the number of students by test mode and year level for both 2018 and 2019.

Table 4: Number of students by test mode and year level, 2018-2019

	Grade 3		Grade 5		Grade 7		Grade 9	
Year	Online	Paper	Online	Paper	Online	Paper	Online	Paper
2018	45,053	236,071	46,305	238,171	44,026	223,121	42,604	202,696
2019	160,844	132,243	159,248	137,384	142,659	145,564	129,518	126,697

All Year 3 students did the NAPLAN writing on paper. In Table 1, the Y3 students who took the other four domains of NAPLAN online are treated as “online” students, and their writing data were combined with Year 5, 7, and 9 online Writing data to fit the IRT models. A different genre was administrated in 2018 and 2019. 2018 writing is a persuasive task and 2019 writing is a narrative task. Each writing submission was marked on ten criteria across all four year levels based on the same marking guide. Most analyses were carried out for the original score categories. In examining the category frequencies, we found that some categories were seldom used, and so for some analyses, these categories were collapsed as listed in Table 5. Note that some criteria, such as 1 and 8, 3 and 6, and 5 and 9, were collapsed in the same way and criteria 2 and 4 were kept intact. The analyses were repeated for collapsed data.

Table 5: Collapsed Categories for Low Frequencies

W01	W02	W03	W04	W05	W06	W07	W08	W09	W10
0-1, 5-6	None	0-1	None	0-1, 4-5	0-1	2-3	0-1, 5-6	0-1, 4-5	0-1-2, 5-6

Table 6 shows the frequency of each category of criteria by year level by test mode for 2018 and 2019 for the original datasets. Table 7 shows the frequency of each category of criteria by year level by test mode for 2018 and 2019 for the collapsed datasets.

Table 6: Frequency of Each Category of Criteria by Year Level by Test Mode for 2018 and 2019

			2018										2019									
			Criteria										Criteria									
Mode	Grade	Score	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Online	3	0	218	4511	559	4720	382	482	14892	682	887	393	424	3482	1111	3411	928	1052	85975	1454	3410	837
		1	3075	17414	4886	14688	1983	6244	23476	4667	8723	2479	6309	33425	12377	29335	7379	19002	72192	13333	40309	7174
		2	24033	21504	23950	23913	39442	36954	6601	27217	24986	15762	79854	114164	84085	112888	133649	134768	2675	97690	87512	50131
		3	17051	1620	15490	1722	3177	1372	84	11761	9835	20466	71546	9759	62716	15156	18501	6013	2	45086	28337	80654
		4	665	4	168	10	69	1	0	718	606	5740	2663	14	552	54	384	9	0	3235	1262	21016
		5	11	0	0	0	0	0	0	8	16	208	48	0	3	0	3	0	0	46	13	1015
		6	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	1	17
	5	0	74	1778	176	1822	133	154	6889	196	545	126	223	1615	428	1604	378	400	38568	540	2087	364
		1	867	9891	1667	7722	646	2520	22351	1789	4908	533	2002	17889	4750	11272	2678	7592	111586	4874	25642	1650
		2	11837	26432	12949	27373	31296	35653	16264	18208	20304	4749	39863	107585	47315	98109	100166	123746	9094	64346	81186	17090
		3	26640	7941	28932	9003	12974	7860	801	21285	18099	21259	94365	31393	98940	46471	49855	27012	0	70465	45478	77093
		4	6383	263	2538	385	1221	118	0	4536	2306	15906	20864	766	7575	1792	5915	498	0	17914	4601	50278
		5	483	0	43	0	35	0	0	282	143	3562	1828	0	240	0	256	0	0	1063	254	12291
		6	21	0	0	0	0	0	0	9	0	170	103	0	0	0	0	0	0	46	0	482
	7	0	55	450	89	455	70	81	3274	102	310	67	190	2021	396	2002	354	366	22839	445	1321	325
		1	298	4897	480	3621	236	1009	15336	856	2726	193	1009	8303	2320	4584	1204	3021	101592	1972	14047	669
		2	4719	20843	5874	21317	18271	26220	22485	10691	15856	1543	15642	70120	20042	57461	58504	87571	18228	33073	63166	5478
		3	21394	16180	28979	16708	20029	15899	2931	20997	20529	12909	70557	57727	93729	70684	64452	49141	0	66974	54797	47374
		4	14499	1656	8078	1925	5068	817	0	10001	4164	17960	45295	4488	24364	7928	16835	2560	0	36057	8684	58225
		5	2801	0	526	0	352	0	0	1306	441	10471	9136	0	1808	0	1310	0	0	3885	644	29039
		6	260	0	0	0	0	0	0	73	0	883	830	0	0	0	0	0	0	253	0	1549
	9	0	80	285	124	298	93	112	1876	125	265	90	273	1947	487	2010	451	470	16859	524	1070	422
		1	231	2492	311	1978	204	625	9178	561	1728	164	787	4066	1302	2039	804	1758	81905	1275	8445	480
		2	2182	12828	2853	13330	9012	17106	23526	6267	10848	644	6894	39352	8332	26665	25511	54967	30754	17530	46853	2108
		3	12694	21037	20959	20753	19132	21388	8024	16693	21025	6660	38067	67592	64656	73242	58368	62863	0	48957	56710	24319
		4	17485	5962	15700	6245	12059	3373	0	14805	7517	13197	52459	16561	46009	25562	38204	9460	0	49606	14674	47108
		5	8291	0	2657	0	2104	0	0	3813	1221	19278	26166	0	8732	0	6180	0	0	10614	1766	50637
		6	1641	0	0	0	0	0	0	340	0	2571	4872	0	0	0	0	0	0	1012	0	4444
Paper	3	0	1889	20930	3540	22036	2669	3305	77506	4331	5436	2708	584	3402	1280	3312	1098	1194	67686	1546	2787	1065
		1	17160	93859	24264	84713	11080	35260	124876	26615	46583	13969	5424	29432	10534	25233	5973	16031	62119	12055	32787	6043
		2	127007	112989	127385	120386	205157	190678	33188	141879	128211	84901	66293	91875	68341	92073	109870	110249	2438	80594	71725	41072

Paper		3	86852	8259	80123	8856	16859	6816	501	59554	52335	104547	57551	7505	51589	11578	14923	4761	0	35274	23704	65825
		4	3098	34	756	80	301	12	0	3635	3391	28714	2349	29	489	47	376	8	0	2731	1222	17317
		5	62	0	3	0	5	0	0	56	115	1212	42	0	10	0	3	0	0	41	18	907
		6	3	0	0	0	0	0	0	1	0	20	0	0	0	0	0	0	0	2	0	14
	5	0	986	8808	1545	9044	1278	1452	28006	1718	2122	1260	330	1397	554	1382	503	538	35794	648	1013	471
		1	4887	42349	8963	38284	3101	11828	114283	9709	17934	3369	1630	13944	3629	9107	1730	5724	92788	3880	14089	1508
		2	61922	142053	67367	144926	164338	189503	92017	95399	110008	29976	32143	90970	37994	83769	82944	107795	8802	52566	68860	14163
		3	141290	43998	150559	44606	64251	35069	3865	108426	95446	106642	82611	30421	88213	41736	46227	22919	0	63067	47909	63950
		4	27398	963	9643	1311	5139	319	0	21918	12038	81405	18921	652	6812	1390	5800	408	0	16172	5280	46847
		5	1658	0	94	0	64	0	0	975	623	14800	1690	0	182	0	180	0	0	1026	233	10013
		6	30	0	0	0	0	0	0	26	0	719	59	0	0	0	0	0	0	25	0	432
	7	0	990	4250	1340	4468	1211	1298	17237	1444	1794	1199	596	1994	823	2003	781	807	30697	850	1074	738
		1	2362	24637	3542	21003	1612	6097	78888	5081	11146	1436	1053	8805	2489	4883	1125	2883	95537	1973	8561	739
		2	29105	107923	34211	110064	95999	143019	113960	55819	82404	11298	15620	72196	20168	60541	58390	89711	19330	30827	59370	5760
		3	111357	79954	147458	80732	99471	69810	13036	104496	103545	63689	71592	58335	96557	71126	65915	49560	0	68112	63903	46136
		4	67316	6357	35009	6854	23515	2897	0	49946	22312	95603	46740	4234	23955	7011	18076	2603	0	38570	11938	63807
		5	11310	0	1561	0	1313	0	0	6063	1920	45873	9223	0	1572	0	1277	0	0	4921	718	26510
		6	681	0	0	0	0	0	0	272	0	4023	740	0	0	0	0	0	0	311	0	1874
	9	0	1456	3400	1736	3498	1599	1691	11657	1815	2009	1591	908	2442	1170	2460	1115	1135	27914	1195	1330	1063
		1	1585	13200	2349	11586	1171	3672	47885	3261	6423	996	929	5653	1929	2797	916	1888	74619	1306	4740	600
		2	14291	67086	16937	70253	48762	93428	113140	32812	54496	4777	8275	44818	10141	33535	31222	58042	24164	17049	40309	2643
		3	68499	98286	109349	97422	93161	92980	30014	79070	99897	31433	42404	62882	70036	71002	57972	58510	0	48098	61730	24251
		4	80977	20724	64686	19937	51118	10925	0	68255	35049	71859	51308	10902	38091	16903	31235	7122	0	47422	17039	52535
		5	31689	0	7639	0	6885	0	0	16232	4822	80993	20009	0	5330	0	4237	0	0	10715	1549	41278
		6	4199	0	0	0	0	0	0	1251	0	11047	2864	0	0	0	0	0	0	912	0	4327

Table 7: Frequency of Each Category of Criteria by Year Level by Test Mode for 2018 and 2019 (Categories Collapsed)

			2018										2019									
			Criteria										Criteria									
Mode	Grade	Score	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Online	3	0	3293	4511	5445	4720	2365	6726	14892	5349	9610	18634	6733	3482	13488	3411	8307	20054	85975	14787	43719	58142
		1	24033	17414	23950	14688	39442	36954	23476	27217	24986	20466	79854	33425	84085	29335	133649	134768	72192	97690	87512	80654
		2	17051	21504	15490	23913	3177	1372	6685	11761	9835	5740	71546	114164	62716	112888	18501	6013	2675	45086	28337	21016
		3	665	1620	168	1722	69	1	0	718	622	213	2663	9759	552	15156	387	9	2	3235	1275	1032
		4	11	4	0	10	0	0	0	8	0	0	48	14	3	54	0	0	0	46	1	0
		5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	941	1778	1843	1822	779	2674	6889	1985	5453	5408	2225	1615	5178	1604	3056	7992	38568	5414	27729	19104
		1	11837	9891	12949	7722	31296	35653	22351	18208	20304	21259	39863	17889	47315	11272	100166	123746	111586	64346	81186	77093
		2	26640	26432	28932	27373	12974	7860	17065	21285	18099	15906	94365	107585	98940	98109	49855	27012	9094	70465	45478	50278
		3	6383	7941	2538	9003	1256	118	0	4536	2449	3732	20864	31393	7575	46471	6171	498	0	17914	4855	12773
		4	504	263	43	385	0	0	0	291	0	0	1931	766	240	1792	0	0	0	1109	0	0
		5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	353	450	569	455	306	1090	3274	958	3036	1803	1199	2021	2716	2002	1558	3387	22839	2417	15368	6472
		1	4719	4897	5874	3621	18271	26220	15336	10691	15856	12909	15642	8303	20042	4584	58504	87571	101592	33073	63166	47374
		2	21394	20843	28979	21317	20029	15899	25416	20997	20529	17960	70557	70120	93729	57461	64452	49141	18228	66974	54797	58225
		3	14499	16180	8078	16708	5420	817	0	10001	4605	11354	45295	57727	24364	70684	18145	2560	0	36057	9328	30588
		4	3061	1656	526	1925	0	0	0	1379	0	0	9966	4488	1808	7928	0	0	0	4138	0	0
		5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	311	285	435	298	297	737	1876	686	1993	898	1060	1947	1789	2010	1255	2228	16859	1799	9515	3010
		1	2182	2492	2853	1978	9012	17106	9178	6267	10848	6660	6894	4066	8332	2039	25511	54967	81905	17530	46853	24319
		2	12694	12828	20959	13330	19132	21388	31550	16693	21025	13197	38067	39352	64656	26665	58368	62863	30754	48957	56710	47108
		3	17485	21037	15700	20753	14163	3373	0	14805	8738	21849	52459	67592	46009	73242	44384	9460	0	49606	16440	55081
		4	9932	5962	2657	6245	0	0	0	4153	0	0	31038	16561	8732	25562	0	0	0	11626	0	0
		5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Paper	3	0	19049	20930	27804	22036	13749	38565	77506	30946	52019	101578	6008	3402	11814	3312	7071	17225	67686	13601	35574	48180
		1	127007	93859	127385	84713	205157	190678	124876	141879	128211	104547	66293	29432	68341	25233	109870	110249	62119	80594	71725	65825
		2	86852	112989	80123	120386	16859	6816	33689	59554	52335	28714	57551	91875	51589	92073	14923	4761	2438	35274	23704	17317

	3	3098	8259	756	8856	306	12	0	3635	3506	1232	2349	7505	489	11578	379	8	0	2731	1240	921
	4	65	34	3	80	0	0	0	57	0	0	42	29	10	47	0	0	0	43	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	5873	8808	10508	9044	4379	13280	28006	11427	20056	34605	1960	1397	4183	1382	2233	6262	35794	4528	15102	16142
	1	61922	42349	67367	38284	164338	189503	114283	95399	110008	106642	32143	13944	37994	9107	82944	107795	92788	52566	68860	63950
	2	141290	142053	150559	144926	64251	35069	95882	108426	95446	81405	82611	90970	88213	83769	46227	22919	8802	63067	47909	46847
	3	27398	43998	9643	44606	5203	319	0	21918	12661	15519	18921	30421	6812	41736	5980	408	0	16172	5513	10445
	4	1688	963	94	1311	0	0	0	1001	0	0	1749	652	182	1390	0	0	0	1051	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	3352	4250	4882	4468	2823	7395	17237	6525	12940	13933	1649	1994	3312	2003	1906	3690	30697	2823	9635	7237
	1	29105	24637	34211	21003	95999	143019	78888	55819	82404	63689	15620	8805	20168	4883	58390	89711	95537	30827	59370	46136
	2	111357	107923	147458	110064	99471	69810	126996	104496	103545	95603	71592	72196	96557	60541	65915	49560	19330	68112	63903	63807
	3	67316	79954	35009	80732	24828	2897	0	49946	24232	49896	46740	58335	23955	71126	19353	2603	0	38570	12656	28384
	4	11991	6357	1561	6854	0	0	0	6335	0	0	9963	4234	1572	7011	0	0	0	5232	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	3041	3400	4085	3498	2770	5363	11657	5076	8432	7364	1837	2442	3099	2460	2031	3023	27914	2501	6070	4306
	1	14291	13200	16937	11586	48762	93428	47885	32812	54496	31433	8275	5653	10141	2797	31222	58042	74619	17049	40309	24251
	2	68499	67086	109349	70253	93161	92980	143154	79070	99897	71859	42404	44818	70036	33535	57972	58510	24164	48098	61730	52535
	3	80977	98286	64686	97422	58003	10925	0	68255	39871	92040	51308	62882	38091	71002	35472	7122	0	47422	18588	45605
	4	35888	20724	7639	19937	0	0	0	17483	0	0	22873	10902	5330	16903	0	0	0	11627	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

3.2. Frequency Distributions

Frequency distributions (FDs) of scores assigned to categories within each criteria, as well as across the total score, can help support or refute inferences about how markers are using the rubrics in practice. These frequency distributions will also reveal which scores are infrequently used by markers. The frequency distributions have been calculated for assessment years 2012-2019.

One pattern visible in the FDs is a tendency towards sparse categories. There are categories in the majority of criteria in every year that are infrequently assigned. Sparse categories may need to be collapsed in analysis, as IRT models require sufficient sample sizes for stable estimation. From a rubric use point of view, sparse categories may occur for a number of reasons. Essays that fit the description actually may be quite infrequent in the data set; the categories may be poorly defined so that scores are classified into clearer adjacent categories; or markers may avoid using the extreme categories of a rubric.

The sparseness in NAPLAN writing data is interesting in its structure. Almost all the sparse categories are those at the top or bottom of the scale, which is a quite common pattern. But for most FDs in the set, there is a tendency towards 'diagonalisation' of the data when viewed across the year levels. An example is shown in Table 8.

Table 8: 2017 frequency distribution for W01 (Audience)

Year	Level	Criteria	% 0	% 1	% 2	% 3	% 4	% 5	% 6	% Missing
2017	Y3	W01	0.356	5.122	45.675	39.750	1.520	0.020	0.002	7.554
2017	Y5	W01	0.153	1.316	18.974	59.136	12.809	0.812	0.021	6.779
2017	Y7	W01	0.240	0.971	10.319	41.867	31.410	7.029	0.597	7.567
2017	Y9	W01	0.357	0.758	5.480	25.152	35.206	18.324	3.603	11.120

The table is highlighted so that cells with less than 1% of the data are shaded in gray, and cells with more than 50% are shown with a heavy box border. The criterion Audience has 7 levels, and broadly speaking, 4 of those score categories are used functionally by the markers for each year level of students. The other levels are quite sparse. However, the specific set of score levels used by markers shifts as the responding students get older. Score category 0 is rarely assigned at any year level. 1 is assigned regularly at Year 3, but not at any other level. Score categories 5 and 6 are very rarely used for Years 3 and 5; category 5 is used at Year 7; and category 6 is used almost exclusively for essays from Year 9. The result is a tendency for the functional levels in the criterion to shift from a block used at the lower end for Year 3 to the higher end at Year 9 - creating a diagonal pattern of data. The diagonalisation is not entirely unexpected given that students' writing skills generally mature as they do. Most would expect older students to write higher-quality texts in Year 9 than in Year 3 as their skills and learning increase.

Another noticeable pattern in the FDs is the clustering of data in specific year level/score category combinations for some criteria. For Vocabulary (W05), the Year 3 students receive a score of 2 more than 70% of the time in every year. The same is true for Cohesion (W06) for Year 3 students; for Year 5 students the percentage in score category 2 in Cohesion is either just below or just above 70% as well.

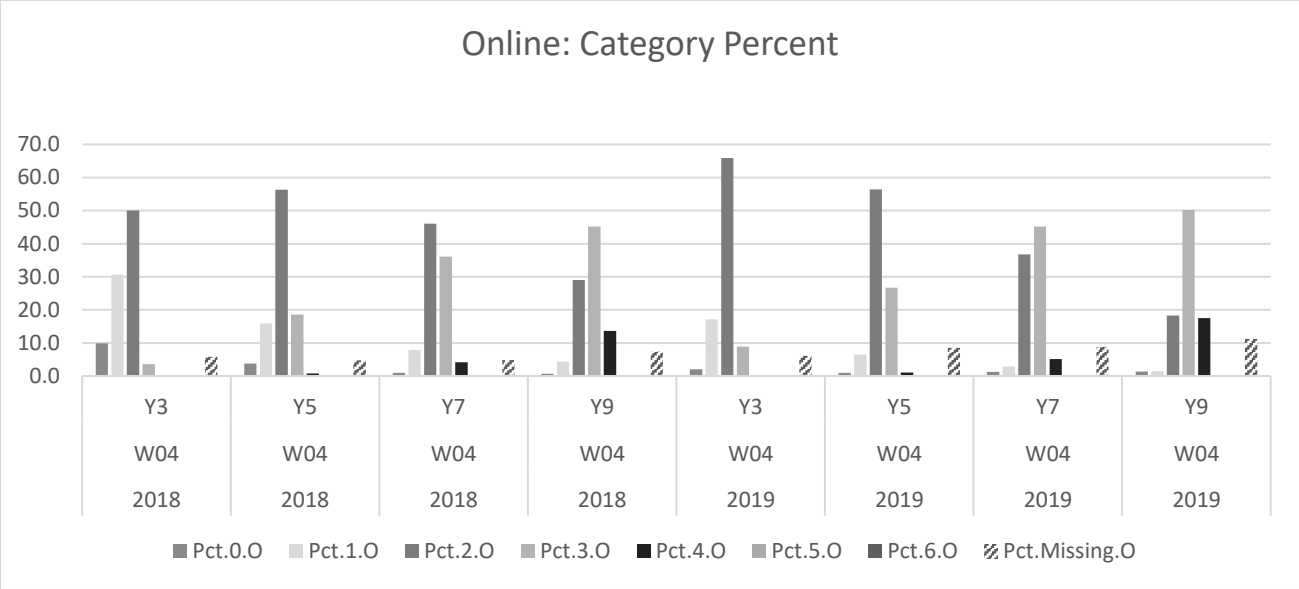
The frequency with which these patterns are seen suggests that different versions of the rubric realistically are in play. NAPLAN moved to using different essay prompts for Year 3/5 and Year 7/9 students based on the idea that the age of the students would have an impact on the type of stimulus each group found engaging. It may be worth considering use of a different rubric at Year 3/5 from the one applied at Year 7/9. This has potential drawbacks in terms of creating a vertical scale like the one currently used to report NAPLAN scores and would have to be carefully investigated before implementation. The patterned use of the categories on the current rubric suggests that the types of essays seen within the age groups are not the same. Construction of different rubrics fit to the two age groupings may allow for more appropriate and detailed description of the classes of responses that appear in each, and might improve marker utilisation of the full range of score categories.

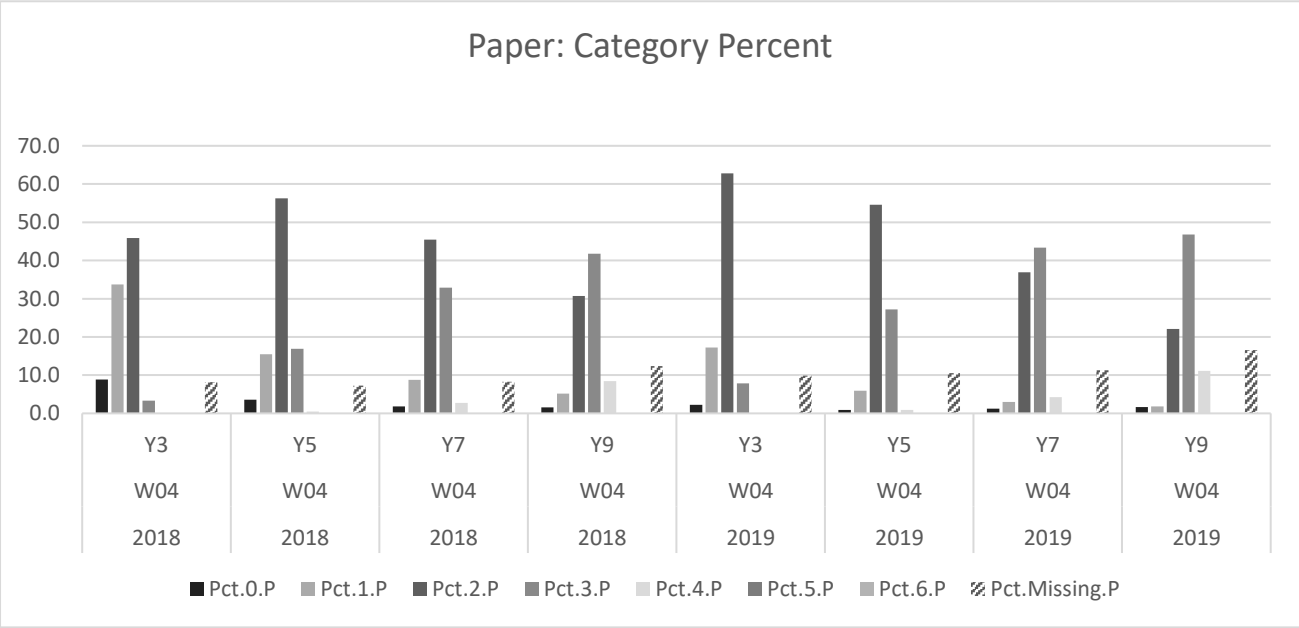
For 2018 and 2019 the FDs were calculated for paper and online responses separately. These were completed to help evaluate whether students writing in different text type or modes (narrative or persuasive, online or paper) are receiving marks with similar distributions. It might be expected that score distributions differ for paper and online scripts. One reason these may differ is that some states assign their best markers to online scripts. There is some evidence that better markers mark online scripts more harshly. Another reason is that online scripts are generally longer and there is some evidence that markers are more inclined to give high scores for longer scripts. Finally, online scripts are generally easier to read than paper scripts. Since the impact of these factors is not unidirectional, it was unclear what the cumulative effect might be.

There is little evidence of differential marking of the NAPLAN writing scripts across text types or mode of response. An example plot is shown in Figure 3. For this display, criterion W04 was selected as it is the most distinct when reading the criteria across text types. For Narrative, the criterion is Character and Setting; for Persuasive, it is Persuasive Devices. It seems possible that if there are differences, they may be most visible when criteria differ across text type.

Although there are small variations in specific score categories, generally the percentage of marks in each rubric level and the overall patterns are quite consistent. This is true for the other criteria as well. There is little evidence to support overall patterns of score differences between online and paper responses.

Figure 3: Percent of marks in each level, online and paper (2018-2019)





3.3. Dependence Indicators

Many common analysis models and techniques assume that the data are 'locally independent', meaning that scores assigned are not related to each other except through the level of latent trait being measured. When there is an apparent relationship that goes beyond the trait, the data are said to be dependent. Data from constructed response items frequently deviate from the assumption of local independence, and for some excellent reasons.

Statistical dependence may be considered as being partitioned into two pieces. One can be described as 'structural dependence'. This local dependence is due to the nature of the task. In NAPLAN writing, as well as many other complex task response types, it is based in the requirement that a marker judge a single performance on multiple criteria which are related to each other. This type of dependence is embedded in the structure of the marking task and it cannot be removed by any effort to refine a rubric, re-train or better train markers, or any other effort that exists outside the closed system of the

task, response, and criteria. This type of dependence is an expected aspect of this type of performance assessment/rubric structure.

The other piece of the statistical dependence is something that potentially can be partly mitigated. The entire process of creating a systematised rubric and training markers in its use is targeted at this reduction. An untrained reader, if asked to evaluate a piece of writing, will tend to assess it on an implied scales of quality, probably ranging broadly from bad to good in most people. This impression is a holistic one, where holistic is used in the sense of matters relating to complete structures rather than with analysis of, treatment of, or subdivision into parts. Holistic writing rubrics impose some organisation onto that judgement, categorising the 'bad to good' into a set of more-specific classes while maintaining the baseline idea of an overall evaluation of quality. Holistic rubrics attempt to increase consistency over what is seen as 'bad' or 'good' across markers by structuring individual's cognitive and continuous conception of writing quality into discrete categories defined by particular characteristics of writing quality. The more that markers can be brought to agree with and use the same definitions of quality, the more consistent—and hopefully accurate—the marking will be. The ultimate goal of marker training is to make it so that, regardless of the specific marker selected to score a piece of writing, the marks assigned to the work will be the same.

Analytic rubrics further extend the idea of structuring holistic judgement into finer-grained and more-specific categories. Analytic marking rubrics take as their premise that writing quality has component parts that can be understood by thinking about how the separate pieces work together to produce the larger, holistic effect. These components ideally would be independent, but in reality they are rarely, if ever, that separable. The components must describe various aspects of the same submission, and that inevitably induces dependence in the data. The extent of the dependence can be mitigated by clear descriptions of components that are designed to be as distinct as possible and that do not intentionally use the same evidence repeatedly.

Rubrics are best constructed so that the set of criteria contains the minimum number of components needed to span the full construct. Each criterion should have the minimum number of score levels so that all responses observed (or reasonably expected to be observed) in the data are suitably categorised and all meaningful differences in performance fall into distinct levels. Score levels within criteria should have language that is as simple as possible while accurately describing the knowledge or skill in that category, and it is preferable that the descriptive language used in each level is as parallel as possible. Component skills belonging to a criterion should be included in the description of all score levels. All of these factors will reduce the cognitive load demanded of the markers using the rubric operationally as well as reducing the degree of local dependence in the data.

3.4. Exploratory Factor Analysis (EFA)

Exploratory factor analysis provides another angle on the complexity and the overlap of criteria in the obtained scored data sets. The NAPLAN rubric currently comprises 10 criteria; there is an underlying assumption that there are sufficient distinctions in performance on each of these that permit them to make a unique contribution to the diagnostic picture of writing resulting from the data. Exploratory factor-analytic approaches can help support or refute the theoretical model of writing that is posited by the current rubric structure. The results may also help diagnose overlapping criteria.

EFA were first fitted to the original 16 datasets. The results of the EFA point to a 1-factor model. Figure 2 shows the scree plot of the EFA analysis on the essay scores of Grade 3 students who took the online test in 2018. The scree plots of the other 15 datasets were similar to Figure 4. The four

methods of eigenvalues, parallel analysis, optimal coordinates, and acceleration factor all showed that one factor is sufficient to explain the variance in the data.

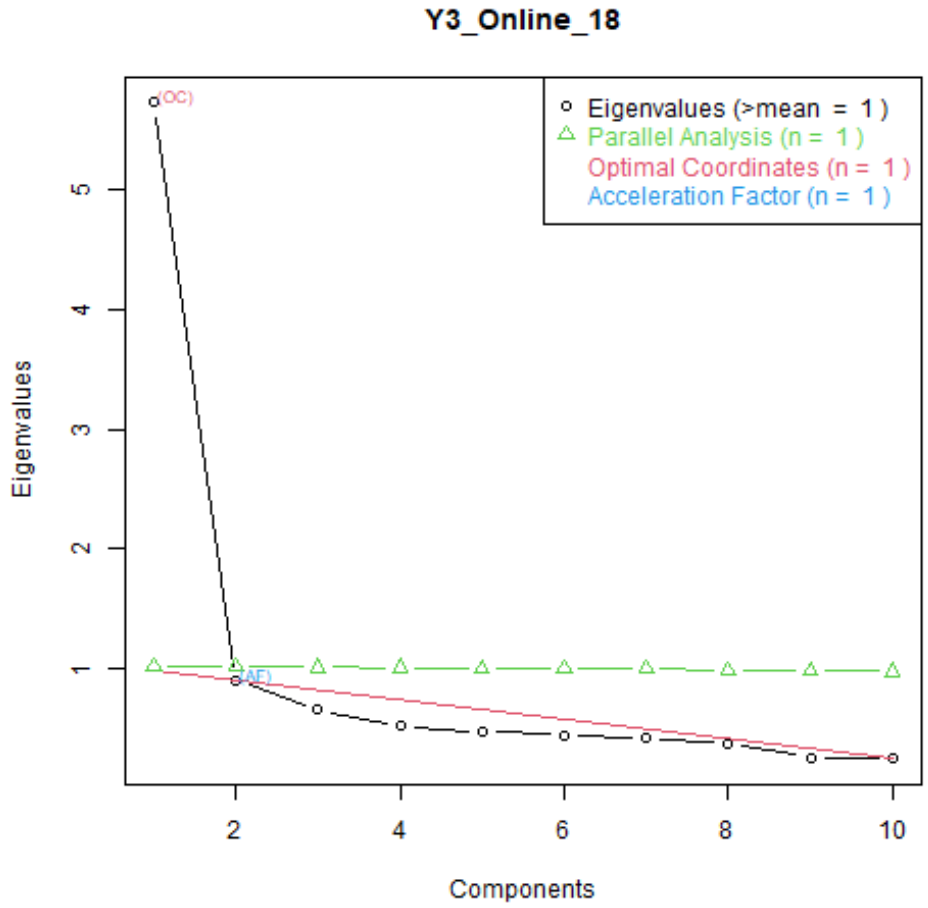


Figure 4: Scree Plot from EFA, online data from Year 3 students, 2018.

There is no evidence for more than one factor from these scree plots. This suggests the assessment is measuring one construct and are potentially influenced by a halo effect. The fact that there is one major factor points to some possible underlying causes. It could be that writing simply is a single unitary trait, and that any subdivision of it will reflect that. If that is the case, then a holistic approach to scoring writing is most appropriate. However, there is substantial evidence that the NAPLAN writing data are statistically dependent. If that dependence can be reduced, a different factor structure, perhaps with more distinct sub-traits, may be observed.

Table 9 shows the loadings of the ten criteria on the factors in the one-factor analyses. Interestingly, Criteria W05 (Vocabulary) and W06 (Cohesion) had higher loadings for the paper form than the online one. Loadings that are 0.70 or above indicate that half or more of the variance in the criterion is accounted for by a single-factor model.

Table 9: Loadings of the 1-Factor Model for the 16 Datasets

	Grade 3				Grade 5				Grade 7				Grade 9			
	Online		Paper		Online		Paper		Online		Paper		Online		Paper	
	18	19	18	19	18	19	18	19	18	19	18	19	18	19	18	19
Criteria	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
1	**0.85**	**0.88**	**0.90**	**0.92**	**0.86**	**0.87**	**0.90**	**0.92**	**0.83**	**0.87**	**0.90**	**0.92**	**0.83**	**0.86**	**0.89**	**0.92**
2	**0.74**	**0.81**	**0.83**	**0.85**	**0.75**	**0.80**	**0.82**	**0.85**	0.69	**0.73**	**0.78**	**0.81**	**0.70**	**0.73**	**0.77**	**0.80**
3	**0.79**	**0.80**	**0.82**	**0.86**	**0.82**	**0.80**	**0.83**	**0.87**	**0.82**	**0.81**	**0.83**	**0.87**	**0.82**	**0.79**	**0.82**	**0.86**
4	**0.74**	**0.80**	**0.83**	**0.86**	**0.75**	**0.79**	**0.83**	**0.86**	**0.72**	**0.77**	**0.80**	**0.83**	**0.74**	**0.76**	**0.80**	**0.84**
5	0.62	0.67	**0.76**	**0.83**	0.64	0.67	**0.78**	**0.84**	0.67	**0.72**	**0.79**	**0.84**	0.68	**0.73**	**0.80**	**0.84**
6	0.67	0.69	**0.75**	**0.79**	0.69	0.69	**0.76**	**0.81**	0.69	0.69	**0.72**	**0.76**	**0.70**	**0.70**	**0.75**	**0.78**
7	0.66	**0.74**	**0.76**	**0.78**	0.67	**0.71**	**0.74**	**0.76**	0.40	0.50	0.52	0.54	0.41	0.45	0.48	0.49
8	**0.74**	**0.72**	**0.75**	**0.78**	**0.75**	**0.74**	**0.78**	**0.81**	**0.75**	**0.75**	**0.76**	**0.78**	**0.75**	**0.76**	**0.78**	**0.81**
9	0.69	0.63	0.62	0.65	0.69	0.66	0.67	**0.70**	0.64	0.61	0.59	0.61	0.65	0.64	0.64	0.65
10	**0.72**	**0.71**	**0.71**	**0.72**	**0.73**	**0.70**	**0.73**	**0.75**	**0.73**	**0.71**	0.69	**0.70**	**0.74**	**0.71**	**0.73**	**0.75**

Table 10: Loadings of the 2-Factor Model for the 16 Datasets

	Grade 3								Grade 5								Grade 7								Grade 9							
	Online				Paper				Online				Paper				Online				Paper				Online				Paper			
	18		19		18		19		18		19		18		19		18		19		18		19		18		19		18		19	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
1	**0.66**	0.52	0.56	**0.68**	**0.76**	0.49	**0.78**	0.50	**0.66**	0.54	**0.68**	0.54	**0.74**	0.51	**0.73**	0.56	0.44	**0.75**	**0.78**	0.44	**0.75**	0.50	**0.76**	0.53	0.43	**0.76**	**0.78**	0.43	**0.76**	0.48	**0.76**	0.52
2	0.29	**0.87**	0.37	**0.77**	**0.73**	0.42	**0.74**	0.43	0.30	**0.85**	**0.76**	0.36	**0.74**	0.40	**0.76**	0.43	0.46	**0.50**	**0.63**	0.38	**0.74**	0.33	**0.77**	0.35	0.46	**0.52**	**0.63**	0.38	**0.73**	0.33	**0.76**	0.35
3	**0.60**	0.50	0.50	**0.62**	**0.71**	0.42	**0.75**	0.43	**0.61**	0.52	**0.65**	0.47	**0.69**	0.46	**0.70**	0.52	0.36	**0.82**	**0.75**	0.38	**0.70**	0.44	**0.72**	0.49	0.37	**0.82**	**0.74**	0.36	**0.72**	0.42	**0.73**	0.47
4	0.37	**0.72**	0.38	**0.74**	**0.73**	0.41	**0.75**	0.42	0.38	**0.71**	**0.73**	0.38	**0.74**	0.41	**0.75**	0.45	**0.51**	0.50	**0.61**	0.46	**0.71**	0.40	**0.75**	0.41	0.52	0.51	**0.61**	0.45	**0.70**	0.41	**0.73**	0.43
5	**0.61**	0.24	**0.55**	0.40	0.58	0.49	**0.67**	0.50	**0.63**	0.26	0.42	**0.54**	**0.56**	0.54	**0.60**	0.59	**0.54**	0.40	0.49	**0.54**	0.56	0.56	0.59	**0.60**	**0.57**	0.39	0.48	**0.55**	0.56	**0.57**	0.59	**0.61**
6	**0.65**	0.27	**0.59**	0.40	0.53	**0.54**	0.56	**0.57**	**0.66**	0.30	0.42	**0.56**	0.53	**0.54**	0.56	**0.59**	**0.58**	0.39	0.45	**0.53**	0.49	**0.54**	0.50	**0.59**	**0.59**	0.40	0.47	**0.52**	0.52	**0.54**	0.53	**0.58**
7	0.42	**0.53**	0.50	**0.55**	**0.61**	0.45	**0.64**	0.45	0.41	**0.55**	**0.55**	0.45	**0.59**	0.43	**0.63**	0.44	**0.38**	0.19	0.31	**0.40**	0.33	**0.40**	0.36	**0.41**	**0.39**	0.20	0.25	**0.40**	0.30	**0.40**	0.31	**0.39**
8	**0.71**	0.32	**0.69**	0.36	0.43	**0.69**	0.45	**0.72**	**0.70**	0.35	0.39	**0.69**	0.45	**0.69**	0.45	**0.72**	**0.65**	0.42	0.43	**0.66**	0.42	**0.69**	0.43	**0.71**	**0.66**	0.42	0.44	**0.66**	0.44	**0.69**	0.45	**0.72**
9	**0.62**	0.35	**0.61**	0.31	0.31	**0.63**	0.34	**0.63**	**0.61**	0.36	0.33	**0.64**	0.34	**0.66**	0.35	**0.66**	**0.63**	0.28	0.26	**0.64**	0.25	**0.64**	0.27	**0.64**	**0.64**	0.29	0.27	**0.67**	0.28	**0.68**	0.29	**0.67**
10	**0.68**	0.31	**0.66**	0.36	0.46	**0.57**	0.50	**0.53**	**0.69**	0.34	0.37	**0.65**	0.45	**0.61**	0.47	**0.61**	**0.64**	0.40	0.42	**0.60**	0.40	**0.60**	0.42	**0.59**	**0.65**	0.40	0.41	**0.63**	0.43	**0.62**	0.45	**0.63**

The exploratory factor analysis results indicate strongly that there is a single underlying factor. However, local dependence in the data may affect the results of such analyses, and its presence may overwhelm the presence of smaller but still meaningful factors. To examine whether there is more than one dimension behind the criteria, a 2-factor analysis was completed. One possible partition of the NAPLAN writing rubric is a separation between the authorial criteria (W01-W06) and the language conventions criteria (W07-W10). If this construction is correct, then the factor loadings patterns should support it. Table 10 shows the loadings on two factors for the original datasets.

Factor loadings revealed fairly regular patterns for Year 7 and 9 students, with one factor containing larger loadings on criteria W01-W04 and the other W05-W10, although criterion W05 (Vocabulary) often is nearly equally weighted on both factors. Scores for Years 3 and 5 students displayed different patterns for the online and paper test forms and the patterns are not entirely consistent across years. For the online form, in 2018, one factor included W01, W03, W05, W06, W08, W09, and W10 and the other one W02, W04, and W07. In 2019, this pattern was not as clear; W01 and W03 were more heavily loaded onto the factor with W02, W04, and W07, although these two criteria loaded relatively strongly on both factors. For the paper tests, one factor had larger loadings for W01-W04 and W07, and the other factor for W08-W10. W05 and W06 (Vocabulary and Cohesion) often were similarly loaded on both factors. This is not true of the 2018

The 2-factor analysis results do not support the division into authorial and language conventions aspects of writing. The pattern of results do suggest that, if these two aspects are posited as an underlying theoretical structure, then criteria W05 and W06 are not clearly aligned with either aspect.

3-factor and 4-factor analyses (results not shown) find that criteria W08-W10 tended to cluster together on one factor and that criteria W01-W04 showed different trends for the paper and online tests. Those four criteria were loaded on one factor for the paper tests; however, for the online tests, criteria W01 and W03 loaded on one factor and W02 and W04 onto another.

Tables 11 and 12 show the loadings of the ten criteria on the factors in the 1-factor and 2-factor analyse on the 16 datasets with some categories collapsed per Table 5. As can be seen, the results are very similar to those from the original data.

3.5. IRT Analyses

3.5.1. One-dimension (1D) Partial Credit Model (PCM)

To further examine local independence among the criteria, a one-dimension partial credit Rasch model analysis was completed for each of the original four datasets: online 2018, online 2019, paper 2018, and paper 2019. Each criterion was treated as one polytomous item. Table 13 shows the weighted fit values for all possible criterion pairs. It seems that criterion W09 (Punctuation) has the strongest local dependence (LD) issues with criteria W06, W07, W08, and W10 (Cohesion, Paragraphing, Sentence Structure and Spelling), shown by the biggest average fit values. This finding is consistent with the results from the EFA, where in the 2-, 3-, or 4-factor models, criterion W09 loaded on the same factor as criteria W08 and W10. With the exception of W06, these criteria are those proposed as the language conventions cluster above and this analysis suggests that these pairwise combinations with W09 are more strongly related.

3.5.2. 1D Generalised Partial Credit Model (GPCM)

We also fit the 1D 2PL generalised partial credit models on the four original datasets to further examine which criteria may have LD issues. Table 14 shows the tau values for the criteria. We have detected six criteria with tau > 2, including Text Structure, Cohesion, Vocabulary, Persuasive Devices/Character and Setting, Ideas, and Audience (W01-W06). Note that these criteria comprise the set posited as the authorial aspects of writing. Sentence Structure (W08) has a tau value that is very close to 2.

Audience (W01) has the largest value of 5.31 and should be examined further, potentially either excluded from the marking rubric or revised to make it more distinguishable from the other criteria. In other analyses, Audience is the criterion score most predictive of total score, so here it may be acting to some extent as an 'overall' quality indicator for the essay. If this is the case, then the large tau values may be a result of similar data-analytic structures as those found in the EFA results described above.

3.5.3. Two-dimension (2D) Generalised Partial Credit Model

The 2D GPCM model was fit on the four collapsed datasets, with dimensions the authorial cluster of criteria W01-W06 (Audience, Text Structure, Ideas, Persuasive Devices/Character and Setting, Vocabulary, and Cohesion) and the language conventions cluster of criteria W07-W10 (Paragraphing, Sentence Structure, Punctuation, and Spelling). Table 15 shows the correlations between the two dimensions for each of the four datasets. All the correlations were greater than 0.9, which means that the rubric failed to distinguish between these dimensions.

Table 11: Loadings of the 1-Factor Model for the 16 Datasets (Categories Collapsed)

	Grade 3				Grade 5				Grade 7				Grade 9			
	Online		Paper		Online		Paper		Online		Paper		Online		Paper	
	18	19	18	19	18	19	18	19	18	19	18	19	18	19	18	19
Criteria	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
1	**0.85**	**0.88**	**0.90**	**0.92**	**0.85**	**0.87**	**0.90**	**0.92**	**0.84**	**0.87**	**0.90**	**0.92**	**0.84**	**0.85**	**0.89**	**0.91**
2	**0.76**	**0.81**	**0.83**	**0.85**	**0.77**	**0.80**	**0.83**	**0.85**	0.68	**0.73**	**0.78**	**0.81**	**0.70**	**0.73**	**0.77**	**0.80**
3	**0.78**	**0.79**	**0.81**	**0.85**	**0.80**	**0.78**	**0.82**	**0.85**	**0.81**	**0.80**	**0.82**	**0.86**	**0.82**	**0.78**	**0.81**	**0.85**
4	**0.75**	**0.8**	**0.83**	**0.86**	**0.76**	**0.80**	**0.83**	**0.86**	**0.72**	**0.76**	**0.80**	**0.84**	**0.73**	**0.76**	**0.80**	**0.84**
5	0.57	0.66	**0.75**	**0.81**	0.58	0.65	**0.76**	**0.82**	0.63	**0.71**	**0.77**	**0.81**	0.63	**0.72**	**0.78**	**0.82**
6	0.62	0.68	**0.74**	**0.78**	0.64	0.66	**0.74**	**0.79**	0.65	0.68	**0.71**	**0.74**	0.66	**0.68**	**0.73**	**0.76**
7	0.68	**0.72**	**0.70**	**0.68**	0.69	**0.71**	**0.69**	0.68	0.40	0.50	0.52	0.54	0.42	0.45	0.48	0.49
8	**0.71**	**0.71**	**0.74**	**0.77**	**0.71**	**0.73**	**0.77**	**0.79**	**0.72**	**0.74**	**0.75**	**0.77**	**0.72**	**0.74**	**0.77**	**0.79**
9	0.66	0.62	0.61	0.63	0.66	0.64	0.65	0.67	0.61	0.59	0.58	0.59	0.61	0.62	0.62	0.62
10	0.62	0.69	0.69	**0.70**	0.63	0.66	**0.70**	**0.72**	0.64	0.68	0.67	0.67	0.64	0.68	0.69	**0.71**

Table 12: Loadings of the 2-Factor Model for the 16 Datasets (Categories Collapsed)

	Grade 3								Grade 5								Grade 7								Grade 9							
	Online				Paper				Online				Paper				Online				Paper				Online				Paper			
	18		19		18		19		18		19		18		19		18		19		18		19		18		19		18		19	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
1	**0.69**	0.48	**0.69**	0.54	**0.77**	0.48	**0.78**	0.49	**0.72**	0.45	**0.71**	0.50	**0.75**	0.49	**0.74**	0.54	**0.77**	0.41	**0.78**	0.44	**0.75**	0.49	**0.75**	0.53	**0.78**	0.40	**0.78**	0.42	**0.76**	0.48	**0.75**	0.52
2	0.30	**0.89**	**0.76**	0.38	**0.72**	0.42	**0.74**	0.43	0.32	**0.91**	**0.75**	0.37	**0.73**	0.41	**0.75**	0.43	**0.53**	0.41	**0.64**	0.37	**0.75**	0.32	**0.78**	0.35	**0.54**	0.42	**0.64**	0.36	**0.74**	0.32	**0.77**	0.34
3	**0.61**	0.46	**0.64**	0.47	**0.71**	0.41	**0.74**	0.43	**0.66**	0.44	**0.68**	0.41	**0.70**	0.43	**0.71**	0.48	**0.81**	0.34	**0.75**	0.36	**0.71**	0.43	**0.70**	0.49	**0.82**	0.34	**0.75**	0.34	**0.72**	0.40	**0.72**	0.46
4	0.41	**0.68**	**0.75**	0.38	**0.74**	0.40	**0.76**	0.42	0.47	**0.63**	**0.73**	0.37	**0.74**	0.41	**0.76**	0.44	**0.53**	0.46	**0.61**	0.45	**0.72**	0.39	**0.75**	0.41	**0.53**	0.49	**0.61**	0.44	**0.70**	0.40	**0.74**	0.43
5	**0.54**	0.24	0.41	**0.52**	**0.58**	0.48	**0.65**	0.48	**0.54**	0.25	0.43	**0.49**	**0.56**	0.51	**0.59**	0.56	0.42	**0.47**	0.49	**0.53**	**0.55**	0.54	0.57	**0.59**	0.40	**0.50**	0.48	**0.54**	0.54	**0.56**	0.56	**0.60**
6	**0.59**	0.26	0.41	**0.56**	**0.53**	0.52	0.55	**0.56**	**0.59**	0.28	0.44	**0.51**	**0.53**	0.52	0.55	**0.57**	0.41	**0.51**	0.45	**0.51**	0.49	**0.53**	0.49	**0.58**	0.40	**0.53**	0.46	**0.50**	0.51	**0.52**	0.52	**0.56**
7	0.47	**0.50**	**0.53**	0.49	**0.56**	0.42	**0.56**	0.38	0.49	0.49	**0.54**	0.46	**0.54**	0.43	**0.56**	0.40	0.20	**0.40**	0.31	**0.40**	0.34	**0.40**	0.35	**0.41**	0.20	**0.41**	0.26	**0.39**	0.30	**0.40**	0.31	**0.39**
8	**0.70**	0.28	0.35	**0.69**	0.43	**0.68**	0.44	**0.73**	**0.69**	0.29	0.39	**0.69**	0.44	**0.69**	0.44	**0.72**	0.42	**0.63**	0.43	**0.65**	0.43	**0.68**	0.42	**0.71**	0.41	**0.64**	0.43	**0.65**	0.43	**0.69**	0.44	**0.72**
9	**0.61**	0.31	0.30	**0.61**	0.30	**0.62**	0.34	**0.60**	**0.60**	0.32	0.31	**0.64**	0.32	**0.66**	0.34	**0.65**	0.26	**0.65**	0.25	**0.64**	0.24	**0.63**	0.25	**0.63**	0.26	**0.65**	0.26	**0.66**	0.26	**0.67**	0.27	**0.65**
10	**0.59**	0.27	0.36	**0.63**	0.45	**0.55**	0.51	0.50	**0.58**	0.28	0.36	**0.60**	0.43	**0.58**	0.47	**0.56**	0.36	**0.58**	0.40	**0.58**	0.39	**0.59**	0.40	**0.57**	0.35	**0.58**	0.38	**0.61**	0.40	**0.60**	0.42	**0.61**

Table 13: Weighted Fit Values for Each Item Pair of the 1-D PCM for the Four Datasets

	Weighted Fit				
Criteria	Online18	Online19	Paper18	Paper19	Average
1,6	0.81	0.80	0.87	0.86	0.84
1,7	0.85	0.86	0.87	0.93	0.88
1,5	0.87	0.86	0.92	0.91	0.89
1,8	0.90	0.87	0.94	0.90	0.90
3,6	0.89	0.89	0.94	0.93	0.91
2,8	0.90	0.92	0.91	0.93	0.92
2,5	0.88	0.93	0.92	0.97	0.92
4,8	0.94	0.90	0.96	0.91	0.93
1,4	0.93	0.90	0.96	0.95	0.94
3,8	0.95	0.91	0.98	0.92	0.94
1,9	0.97	0.94	0.97	0.92	0.95
2,6	0.90	0.97	0.93	1.01	0.95
3,5	0.96	0.92	1.01	0.96	0.96
1,2	0.90	0.97	0.93	1.05	0.96
4,6	0.94	0.95	0.98	0.99	0.96
2,10	0.94	0.97	0.97	0.99	0.97
4,10	0.96	0.97	0.99	0.97	0.97
4,5	0.96	0.97	0.98	1.01	0.98
3,7	0.95	0.96	0.98	1.04	0.98
1,10	0.98	0.94	1.04	0.97	0.98
5,6	0.96	0.96	1.03	1.00	0.99
3,4	1.00	0.95	1.01	1.00	0.99
5,8	1.00	0.95	1.04	0.99	1.00
3,10	1.01	0.97	1.07	0.99	1.01
2,3	0.96	1.02	0.98	1.10	1.01
5,7	0.97	1.02	0.99	1.09	1.02
4,7	0.98	1.02	1.01	1.08	1.02
1,3	0.97	1.01	1.05	1.09	1.03
3,9	1.07	1.03	1.05	0.99	1.04
4,9	1.08	1.07	1.06	1.01	1.05
7,8	1.04	1.09	1.05	1.12	1.08
2,9	1.10	1.12	1.07	1.04	1.08
6,10	1.11	1.04	1.15	1.06	1.09
6,8	1.14	1.06	1.14	1.06	1.10
6,7	1.02	1.12	1.06	1.20	1.10
7,10	1.05	1.18	1.08	1.23	1.13
5,9	1.16	1.14	1.15	1.10	1.14
2,7	1.14	1.11	1.17	1.20	1.15
8,10	1.19	1.12	1.22	1.13	1.17
5,10	1.19	1.10	1.26	1.14	1.17
2,4	1.16	1.20	1.19	1.22	1.19
6,9	1.30	1.28	1.25	1.17	**1.25**
9,10	1.42	1.37	1.44	1.33	**1.39**
7,9	1.29	1.56	1.25	1.53	**1.41**
8,9	1.51	1.39	1.44	1.30	**1.41**

Table 14: Tau Values of the 1-D 2PL GPCM for Each Criterion of the Four Datasets

	Tau values				
	Online18	Online19	Paper18	Paper19	Average
Punctuation	1.04	1.03	1.23	1.24	1.13
Paragraphing	1.98	1.18	1.93	0.92	1.50
Spelling	1.45	1.65	1.54	1.69	1.58
Sentence Structure	1.70	2.13	1.92	2.20	1.99
Text Structure	2.66	2.47	2.70	2.25	**2.52**
Cohesion	2.61	2.67	2.67	2.73	**2.67**
Vocabulary	2.55	2.98	2.67	2.90	**2.78**
Persuasive Devices	2.70	2.99	2.68	2.86	**2.81**
Ideas	3.25	4.24	3.41	3.64	**3.63**
Audience	5.06	6.26	4.91	5.01	**5.31**

Table 15: Dimension Correlation of the 2-D GPCM for the Four Datasets (Categories Collapsed)

Correlation Between D1 and D2			
Online18	Online19	Paper18	Paper19
0.96	0.94	0.96	0.93

For each of the analytic approaches described above, there is much greater detail that can be explored. However, there is a clear thread of common findings throughout them all. The observed data from the NAPLAN writing assessment consistently appear to have strong evidence of local dependence. Collapsing categories within a criterion did not appear to have much if any effect on this finding. While there is not an expectation that this dependence can be removed from the data given the structure of the task, steps to reduce it are suggested in the other sections of this paper. As changes are implemented, these analyses can be re-run to evaluate the effect and determine whether the dependence has been reduced substantially.

4. Overall Marking Design

4.1. Marker Training

Effective marker training is a requirement of high-quality marking. This obvious statement is left open to interpretation because “effective” is not clearly defined in the research literature. There are industry rules-of-thumb as well as a developing set of guidance documents. NAPLAN’s marker training will be compared with these, as well as with approaches used in other national and international programs. Differences will be noted for consideration.

NAPLAN’s marker training appears to be generally effective. Inter-rater agreement is acceptably high. Score reliability is quite high, although the presence of local dependence in the data has the effect of inflating these statistics. There are modifications to the NAPLAN writing marker training that could be implemented, which reasonably would be expected to improve the quality and consistency of the marking.

4.1.1. Move marker training online

NAPLAN writing marking involves hundreds of markers nationally, with one or more marking centres in each state. The training of the markers is conducted using a standard set of materials. The leaders of the training in each state are trained using a train-the-trainer model and can be expected to replicate the initial training with reasonable fidelity. It has been shown that individual variation introduced by having multiple leaders of training – and of marking – can have an influence of the quality, consistency, and tendency of marking. Different trainers will emphasise slightly difference aspects of the rubric. They will phrase the verbal components of the training slightly differently. It is possible that some aspects of the training are left unspecified at initial training, so each trainer will have to develop their own understanding of the materials, sample scripts, and cited evidence. This variation is small but can be meaningful. Creating a single online training experience for markers will eliminate this as a source of variance.

If training is moved online, all markers will have exactly the same experience; it is completely standardised. Face-to-face training is almost always more popular with markers than online training, so this change will result in complaints. However, it has been shown that online training is equally effective to marginally more effective on the resulting marking quality. Use of asynchronous, online training also broadens the pool of potential markers. If marking opportunities require that people be able to travel to a site, that eliminates people who have difficulties with travel (physical limitations or family commitments, for example) from the possible pool. Also, by broadening the pool of markers, ACARA would be addressing any equity issues within the cohort of markers, with the potential for a more culturally and socially diverse cohort of markers made possible.

Another advantage of online training is that it is accessible 24/7. The window permitted for marker training can be constrained so that marking scheduling and other needs can be met, but during that period, if a potential marker wants to review materials at 3 a.m., they can do so without issue. And as we have seen this year, if a pandemic occurs, everyone can still train and mark online at home in isolation exactly as planned – assuming student assessments have been administered and there are responses to mark.

Online marker training requires a fairly substantial initial investment of time, funding, and expertise. A system must be developed or licensed. Most commercial content management systems (CMS) are well able to support marker training content. These systems support interactive content, videos of trainers speaking about the materials, formative assessment opportunities, and other assets that

support effective training. In addition, there are specialist suppliers on the market who combine training and marking on a single platform. The advantage of a single platform for training and marking is simplified access to training materials at any time for the markers, so if they wish to refresh themselves on a specific point or review the samples around specific levels on criteria when assigning a mark, it is simple to do so. In some systems, the exemplar cases are linked directly into the score-assignment part of the pages, making access even simpler. Marker activity can be tracked, so that if certain aspects of the training or specific examples are accessed frequently, targeted supplemental activities or Q&A support could be provided across the marker pool. Currently-available communication platforms (*e.g.* Slack) can be used as a supplement in the online marking experience, whereby interactive, real-time communication between lead markers and markers can take place throughout the marking window. This enhances both the confidence of the marker and the reliability of the marking.

4.1.2. Re-certify all markers every year

All markers should be assessed for accuracy before beginning marking for every cycle. Even very experienced markers can drift off-rubric unaware and become subjective in their marking. A number of assessment programs worldwide require that all markers attend training every cycle, especially if the cycles are separated by an extended period of time. If training is moved online, markers can easily select the most challenging parts of the rubric to read, review examples of criteria that are subtly different, or read any new Q&A posted since last cycle. But all markers, new and experienced, should be required to pass the same skills assessment before live marking (as well as QA measures during live marking – see section 4.4 for more details).

It is generally accepted practice to allow markers more than one attempt at accuracy certification within a single marking cycle. Two attempts commonly are permitted, allowing for a marker having an 'off day' or for some unusual interaction with the specific scripts in the certification set attempted. Operational research indicates that markers who pass certification on the second attempt rather than the first are equally accurate overall. Assuming that the marker certification assessment is given online at the end of marker training, the scoring system can manage the necessary logistics of multiple attempts automatically. ACARA provide pre-marked scripts (Online Practice or Qualification scripts) to the Test Administration Authorities (TAAs) that may be used in this manner, but it is not currently required.

Other process controls can be implemented by most well-designed marking software systems. Examples include:

- Requiring a minimum period of review, such as 24 hours, after an unsuccessful certification attempt. This limits the probability of markers using their second attempt immediately after the first, as they may be upset that they have not been successful and this may impact marking performance.
- Using a stratified random sampling algorithm to build certification forms from a pool of pre-scored responses. Stratification generally can be built around strata selected by the program; in NAPLAN's case, year level, essay length, or online/paper response mode might be variables of interest. This approach minimises the likelihood of inappropriate access to certification script scores from certified markers. If this approach is implemented, it is recommended that the pool be at least 2.5x the length of the certification set to increase the number of unique forms that can be assembled (e.g. if the certification assessment comprises 4 scripts, there should be at least 10 scripts in the pool). To account

for the possibility of an interrupted certification attempt, sufficient scripts can be included in the pool to support a third attempt if desired.

4.2. Sample Selection and Exemplar Sets

Related to marker training design and expectations is selection and size of the exemplar set that makes concrete the criteria and categories being measured. As with marker training, there is a small but emerging set of guidance and operational standards in the assessment industry.

Sample scripts have numerous uses in human marking, including marker training and marker skills evaluation. Sample scripts illustrate the trait levels across the criteria such as benchmarks and rangefinders, practice sets, warm-up scripts, certification assessment of marker accuracy (see Section 4.1 for more details), and ongoing assessment of marker accuracy in processes like calibration and control scripts (aka validity scoring). The mechanics and selection criteria for these procedures will be described in Section 4.4 below. However, all of them require scripts pre-marked by experts that serve as a standard of accuracy in marking. NAPLAN marking centres currently are supplied with pre-marked exemplars provided for a variety of uses.

4.2.1. Feedback and Correct Marks

Provision of feedback is a key factor in differentiating uses and types of QA cases. Correct scores *are* provided for QA cases used as part of marker training, and each should have a rationale explaining the specific evidence from the response that aligns with the rubric description of the assigned level. Rationales that include an explanation of why the example was not assigned the adjacent category or categories are especially valuable to trainee and novice markers who may have difficulty distinguishing categories. This means that for an exemplar level 3, the rationale should explain why the 3 was assigned as well as why a 2 or a 4 was not assigned. Rationales should always refer explicitly back to the language of the rubric or criterion with evidence aligned to it. It is helpful to markers, as well as the experts creating them, to use a consistent format or template for mark rationales. These training cases are not used for formal or consequential evaluation of marker skills, although practise scripts may be used formatively by trainee markers as they improve their skills. Training scripts are provided to the TAAs for NAPLAN marker training.

Correct scores *are not* typically given for QA cases used for formal assessment of marker accuracy. The primary purpose of assessing marker accuracy is achieved only when markers apply the rubric to score the case. This can be subverted by at least two sources of information: marker memory or records of a previously-provided correct score. Either may substitute for active application of the criteria. Note that the fact a marker may identify a response as one they have marked before is acceptable. It is only of concern if either the marker remembers the score(s) assigned in the prior scoring and re-assigns them from memory, or if the marker has recorded scores for the response and re-uses those. In these situation, recall of specific marks is replacing active application of the rubric.

In reality, marker memory of the precise scores assigned to QA cases is rarely a significant issue. Operational use has provided quite a bit of evidence that markers have a shorter and less accurate memory for details of the responses than they believe. Markers will vigorously argue that they have scored a particular response more than once when there is clear system evidence that they have not done so. This dispute is frequently accompanied by scoring-system evidence that they actually may have marked QA cases repeatedly without recognising them.

Markers may indeed encounter a QA case more than once during live scoring. In sessions with long active windows, it can be prohibitive in terms of resources, cost and time to create a set of QA cases sufficiently large to ensure that no marker sees any single example on a skills assessment more than once. And as long as a marker reads the response and marks it against the rubric—not from memory—then a QA case serves its purpose in verifying the accuracy of marking. Thus is it common practice to recycle marking accuracy assessment cases during the scoring window. If managed carefully, this method can produce both accurate data and considerable savings. As long as the QA cases are separated by a reasonable interval of time ('reasonable' varies based on the complexity and uniqueness of the response type), markers are unlikely to recall that they have previously seen a specific response, recall the marks they assigned, and apply those memorised marks.

The other sources of external data is feedback on marking performance, marker training, or possibly unauthorised access to correct scores. If markers are told what the correct scores are on a visible skills assessment as part of feedback from a marking leader, they may make note of that information. Should they see the response again (assuming they recall it correctly), they can refer to their notes and assign the recorded marks, thus avoiding a meaningful evaluation of their skills. Markers may also share a description of cases and correct marks for them, intentionally or inadvertently, with other markers who have not yet completed those assessment cases.

Provision of no feedback on marker accuracy assessments, particularly those with stakes for initial or ongoing employment, is inevitably unpopular, and with good reason – markers want to know what they got wrong if they are unsuccessful. Feedback may be provided on marking, either in person or automatically, but it should not expose the 'correct' score for the QA cases. Once this information is known, those QA cases may no longer be viable for assessing marker skill.

4.2.2. Selection Criteria

The criteria for selecting QA cases varies. When selecting exemplars used across all criteria, the following standards are recommended:

- For use in higher stakes contexts such as qualification or continuation of employment (initial **certification** and ongoing **calibration**), the exemplar cases should be clear examples, with evidence that aligns to only one score level for each criterion. The goal in these assessments is not to be tricky or confusing and the samples should not be at the borderline of the score category. Well-trained markers should not struggle overmuch on these QA cases.
- **Control** cases can be less clear on one or a few criteria, as they are not typically used as employment dismissal evidence, except in aggregate. A marker with a cumulative record of repeated inaccuracy on control scripts may be monitored, counselled, re-trained, or terminated, depending on program regulations.
- **Practice sets** can be less clear because the rationale for the marks assigned by the expert is generally provided for the marker's review. This ability to explain unusual features or clarify why one score level and not the adjacent one was assigned supports markers in improving their skills. If correct scores are provided with either a rationale or discussion with marking leadership, **warm-up scripts** may have similar characteristics to practise ones.

In marker training, sample scripts may be chosen so that they exemplify a single score level on one criterion, or possibly for a specific pattern of scores across a subset of criteria. In these cases, the following standards are recommended:

- **Benchmarks** should be clear and unambiguous examples of a single level of a criterion.

- **Rangefinders** are the ‘edge cases’. These are examples that are nearly, but not quite, in the adjacent score category. They are often described as ‘a high 2’ or ‘a low 5’. These examples are difficult to locate and to gain consensus from experts on, and they require very clear and specific rationales about the final score category assigned. But if they can be found, they are extremely valuable in marker training.

4.3. Bias Factors

A third, also related, component of the marking design is biasing factors. It is improbable, if not impossible, to remove bias from human scoring. It is possible to reduce it. All significant biasing factors in the response set should be addressed in marker training. That process begins with defining the most probable biasing factors in the specific assessment program. Once these are clarified, procedures for addressing them can be developed, either through design controls or through explicit marker training.

Marking bias is often framed as an overall tendency where markers assign scores that are, for example, too high (leniency), too low (harshness), or not variable enough (central tendency). Assuming reasonably effective marker training, these overall tendencies are infrequently observed during live marking. If these effects are observed across the full data set, it may be possible to correct the effect statistically in some part.

Bias tends to show up most commonly as an interaction between a subset of responses with some characteristic that results in an inappropriate mark assignment. Marker bias of this type can be triggered by numerous factors. Essay scores such as those in NAPLAN writing may be influenced by response mode (keyed versus hand-written—as long as paper-based testing is supported), writing style, essay length, use of complex vocabulary, and grammar and typographical errors (when these not part of the criterion being measured). Although it is generally thought of as negatively impacting scores, bias can run in either direction. Markers may prefer hand-written essays and so score them more leniently, or react strongly to poor handwriting and mark those more harshly. Each person has their own personal set of biasing factors so the specific subset of cases with potentially biased marks effectively is unique to each individual marker. This makes detecting this type of bias, much less correcting it statistically, nearly impossible. Control of interaction bias is strongly dependent on training markers to be aware of their preferences and reactions so they can limit impact on their work.

4.3.1. Bias factors in NAPLAN Writing

There are specific biasing factors that are relevant to the current NAPLAN marking design.

- Personal knowledge of the examinee has been shown to bias marking. Respondents should remain anonymous to markers whenever feasible. It has also been shown in some studies that investment in the same educational system as respondents – such as the marker being an employee of the same jurisdiction – can influence assigned scores. NAPLAN’s marking design where writing is scored within the state of origin may trigger this bias. A more neutral design should be considered. It would be possible to maintain marking proportional to the number of responses from a state, but to re-distribute the essays so that no state marks its own students’ responses. This approach is relatively simple to implement in an online marking system and possible, although logistically more complex, in paper marking.
- Bias may compound across items if the submission comprises multiple responses, a portfolio of work, or multiple criteria marked against an individual response when scored by a single marker. This effect can significantly disadvantage individual candidates. Scoring multiple or

portfolio submissions item-by-item whenever possible will mitigate this issue, and inclusion of scores from different markers within a response set tends to increase test-score reliability. When responses such as essays are scored using multiple criteria, the effect is more difficult to counter. It is generally not practical to have each criterion scored by a different marker, as the increase in time and resources required to do so is very large. The main mitigation is thorough and explicit bias-reduction training for markers. Inclusion of responses with common bias triggers in the training and QA samples (*e.g.* an essay that is a long but poor piece of writing would tend to be over-scored by markers with a preference for length; one that is succinct but strong would be under-scored by the same group) may support detection of certain types of marker bias, and provide an indication of markers who may need refresher bias and/or marking training.

- Raters who work unusually rapidly or frequently may score a disproportionate number of the total responses in the response pool. If such a rater also displays biased scoring, the whole pool of candidate scores can be distorted by the influence of a single or a few raters. This effect may be exacerbated by paying human raters by the piece rather than by the hour, which can lead to rushed scoring. The pool influence of individual markers can be limited by capping the amount of time markers are permitted to work or the number of responses they are permitted to mark. This can be managed effectively in an online marking system.
- Human markers get tired. The marked data set may display fatigue effects, and these may be more pronounced as responses become more cognitively demanding to evaluate. If remote marking is implemented and markers are permitted to set their own schedules, they may choose to labour for long hours while there are still responses available to score to maximise their earnings. Fatigue effects, like pool influence, can be reduced by imposing limits on maximum hours for raters, within a single working shift as well as across the days or weeks of the scoring window. A degree of cognitive reset may be activated by rotation of the year levels within a marker. This is done in NAPLAN writing marking, where groups of scripts are randomly allocated with the intent that consecutive groups of scripts assigned to a marker be selected from different year groups. As seen in the frequency distributions, there is evidence that students of different ages respond differently and the resulting marks are concentrated in different parts of the scoring scale.

4.4. Quality Assurance Measures

There are numerous quality assurance measures that can be built into human marking designs. Many are greatly facilitated by online software systems that implement such measures automatically. While these measures can be built and executed manually, the cost of such efforts in time and human capital is very high. Quality assurance measures vary in their target of inference, frequency, cost of implementation, and type of risk addressed. Different sets or frequency of measures should be chosen in line with the stakes of the assessment.

What is typically measured in quality assurance designs for human marking is consistency: that scores assigned by two or more raters to the same submission are the same. Consistency is a problematic element of human scoring quality. Agreement is desirable, but raters can agree when they both assign the same incorrect score level. Disagreement is relatively uninformative as well: it is possible that one rater is accurate and the other is not, with no way to determine which is which; or it may be that both raters are inaccurate and the accurate score level is one that neither assigned.

Measures of marker consistency are nonetheless popular, as they are generally simple to calculate and easily explained to and understood by most stakeholders.

Probably the most commonly-used metrics for evaluating marking quality calculate consistency. Measures include percent exact or adjacent agreement, various types of correlation, kappa, tau, and others. The resulting values are commonly referred to as 'inter-rater reliability (IRR)' although agreement measures technically are not part of the class of statistical reliability metrics.

4.4.1. Multi-scored data and disagreed marks

Agreement measures require collection of multi-scored data from a process where more than one score is assigned to the same response by different raters within a single scoring window. These scores should be assigned blindly, in that the additional marker(s) should not know the score level assigned by the previous one(s). Responses can be fully or partially multi-scored, depending on factors like the stakes of the assessment and intended use of the data. If multi-scored data is collected, a standard for defining scores as discrepant should be included in the scoring design. Programs with multiple criteria like NAPLAN increase complexity around decisions about tolerance for marker disagreement. Since each response receives a number of assigned scores, decisions must be made whether disagreement is measured against each criterion individually, at the total-score level, or some combination.

NAPLAN has definitions for discrepant marking against control scripts in the national marking protocol at the total and criterion level. The standard for discrepancy in terms of number of score points for each criterion appears to be the same, even though the criteria range from three to seven score levels. Consideration could be given to setting a discrepancy standard using a different metric instead such as a percentage of the available range or standardised mean difference to trigger actions such as marker monitoring, counselling, re-training, or removal. For example, using 10% of the available range as the standard would mean that for a 3-level criterion scores must agree exactly; a 6-level criterion could have scores no more than 1 point apart; and the total score could be 5 points apart before they are considered discrepant.

If multi-marking is used operationally, there will be cases in which assigned scores disagree. Decisions about which score disagreements, if any, are to be resolved should be made as part of the marking design process, and a process for resolving these scores determined. Maximising unique pairings of raters in multi-scored data designs helps make discrepant marking visible, and most online systems have controls for marker/response pairings to support this. Decisions must be taken about how and if multi-marked scores are to be used or reported operationally; these choices are not without controversy in the field.

4.4.2. Accuracy in marking quality assurance

The conceptual target when designing quality assurance measures for human scoring is actually accuracy: that the score or category assigned by the rater is 'correct' i.e. that it is an unbiased evaluation of the construct being assessed and is assigned to the appropriate level of the rubric. As constructed-response items such as essays generally lack an objective answer, a standard set of scored cases must be developed, based on expert judgement or theory. This standard becomes the criterion measure for a 'correct' categorisation of a response against which rater scores can be evaluated for accuracy. NAPLAN has a large compendium of such expert-scored essays that is used in quality control of the marking.

The process of selecting quality control cases and generating scores for use in accuracy evaluation goes by numerous names: master coding, sample selection, control cases, and expert scoring among others. This process is more complex and challenging than evaluating consistency and so is less commonly used. The unfortunate reality is that a substantial number of assessment programs measure consistency and describe it as accuracy.

One desirable feature of focusing marking quality assurance efforts on accuracy is that markers who are accurate will also be consistent with each other. It is straightforward to measure both with the same multi-scored data set if it is appropriately structured. Note that the QA processes and measures described in this section typically have marker accuracy as the target of inference.

The description below is intended as an outline of an idealised case and not as a description of any specific assessment program's approach. These steps should be modified to align with to the program's needs, stakes, and operational processes.

General steps and QA processes recommended for marker training are listed below.

1. Markers are given access to online training for a fixed window of time. The length of the window is generally set once a reasonable estimate of the length of training is obtained.
2. Training typically begins with a review of the assessment program, items and response types, and an overview of the full rubric. Bias training should be included in the early stages of marker training as well.
3. Following the introductory material, a guided and detailed explanation of each criterion and level is provided. For each score level, the benchmark and rangefinder cases with their rationales are presented to make the rubric description concrete.
    a. Where possible, two benchmark cases should be provided within a score level to illustrate examples that vary in style, length, or other factors yet remain aligned to the 'centre' of a score level.
    b. Rangefinders exist, in theory at least, at the top and bottom of every score category except the highest one (where there is no 'high' rangefinder) and the lowest (where there is no 'low' rangefinder). Rangefinders are particularly desirable for training markers when pairs of score categories appear close together and it is difficult to distinguish between them.
4. After the main marker training on criteria and levels have been reviewed, markers should be given access to practise sets of scripts. Trainee markers score the practice scripts, and their assigned marks are compared to the pre-assigned expert marks. Trainee marker accuracy is calculated, and rationales for the expert scores are provided for all responses—including those scored correctly. It is possible for a trainee marker to assign the correct score level using the wrong evidence, so the expert rationales are essential in correcting those errors. Trainee markers generally prefer that there be several sets of practice marking available, and it is not required that markers complete practice sets. Experience suggests that markers who skip practice sets entirely are substantially less successful in certification.

Once trainee markers have completed training and practised their skills, the marking QA procedures are activated.

5. Certification of markers occurs at the end of training and before markers are admitted to live marking. Given that markers proceed through online training at varying rates, certification generally is opened at the same time or shortly after training is opened. Markers are aware of the assessment, receive a report of the results, and may re-take if necessary.

6. Warm-up scripts serve the same role as practice scripts, differing in the timing. Practice scripts are offered for markers whenever they begin a stage in marking, such as first thing in the morning at system logon or at the time of shifting topics, genres, or changing year levels in marking, as appropriate.
7. Calibration of markers occurs during the live marking window. Markers are aware of the assessment, receive a report of the results, and may re-take if necessary. Calibration usually requires scoring fewer cases on a single assessment than certification. Calibration can be administered after a completed number of live scripts marked (*e.g.* after every 25 live scripts) or on a time schedule (*e.g.* every morning, every 3 days, after 48 hours off-system, etc.)
8. Control cases are allotted during the live marking window. Administration of these cases should be blind to the markers, seeded into live marking at a fixed frequency. Markers do not receive results of this assessment. As with calibration, control scripts can be administered on a periodic schedule. Accuracy results are visible to marking leadership and administrators in the software system.

Operational research suggests that markers will perform differently on visible (certification, calibration) and blind (control scripts) accuracy assessments. Different standards of accuracy may be required for each type.

In comparing the current NAPLAN writing marking procedures to the description above, there are stages that are aligned and steps that differ, based on the national marking protocols.

1. NAPLAN writing marker training is not currently online. It is conducted for a fixed period of time. While this limits access to those markers who can attend on-site, marker training is based on centralised materials and national training to increase consistency across states and territories.
2. Bias training is not currently part of the standard marker training, although it could be added.
3. Markers are provided with extensive sets of pre-scored scripts required for use in training as well as other functions. These training scripts have associated rationales provided to explicate the alignment of the evidence from the essay with the assigned score level. They are not typically classified as benchmark or rangefinder types, although there are some indicated as ‘low’ or ‘high’ examples in the rationales. An explanation of the reasons a response was not categorised into the adjacent score level is occasionally indicated in this set, although not consistently. The exemplars are expert-scored at a consensus marking meeting, with revision, refinement and consultation to assure agreement.
4. Pre-marked scripts are also supplied to TAAs for use as practice scripts, and these are required for anyone marking student submissions in live scoring. These scripts may or may not have automated feedback provided on submission of marks.
5. The pre-marked set of scripts as provided in 4 may be used to certify whether markers are at a suitable level of accuracy to commence marking once they complete training and mark practice scripts. This is not a required step. It may useful to require formal certification for all markers to assure that they are sufficiently accurate to proceed to live marking.
6. An additional set of ‘other’ pre-marked scripts is provided for uses including warm-up scripts, at the discretion of the TAA. These may be supplemented by scripts selected from the operational marking if the centre marking leadership team so chooses. It may be considered whether standardising this process across centres would reduce construct-irrelevant variance in marking.
7. Calibration is not formally used in NAPLAN writing marking.

8. Control scripts are required as part of NAPLAN writing marking. One script, common across all centres marking on that day, is delivered daily to all markers. A more frequent check might be considered, as once-daily may provide limited data and insight, especially for centres who mark for shorter time windows.

NAPLAN implements another quality verification process, check marking. A member of the centre marking leadership systematically checks scripts for marker errors. It is expected that every marker's work will be checked daily at a minimum rate of 10% through this process. Check marking can be blind (i.e. the leader does not view the marker's assigned score for an essay, scoring it independently before evaluating differences), but generally is not as the scored data is used as a pointer towards issues or deviation from prescribed practices. Check marking is triggered and/or guided by indicators such as marker rates of work, unusually high or low scores in aggregate, repeated inaccuracy on one or a few criteria on the control scripts. This process likely varies across marking centres, leadership teams, and initiating causes. Check marking also is used to examine scripts for unexpected commonalities that may be attributed to cheating, memorisation, or pre-practised text in essays.

4.5. Automated Systems

Automated essay scoring (AES) systems encompass a range of levels of complexity and sophistication. Depending on other design factors, such systems could be introduced into NAPLAN in a variety of places and perform different functions. There is a substantial research literature on use of AES for responses across a large range of item types, response lengths, content areas, and stakes of judgements based on the scores. Included in that body of work are previous studies demonstrating that the statistical and technical qualities of AES use in NAPLAN are sound. The technical quality of such scoring increases consistently over time, a feature that is likely to continue as AES are developed for shorter responses and to evaluate features beyond writing quality such as the accuracy content statements. Inclusion of artificial intelligence and natural language processing capabilities in AES will continue to push the boundaries of what computers can effectively score.

In the rush of technological development, human factors may get sidelined or discounted. However, stakeholder management and reaction are key considerations for the introduction of AES into any assessment program, and especially a high-profile one like NAPLAN. As NAPLAN moves fully online, it will be technologically easier and easier to implement the software needed. That is not to say that specific research studies supporting specific AES uses in NAPLAN can be omitted—they cannot. But public perception, policy positions, and educator expectations will all require management that is just as careful as that needed to assure that the AES technical specifications and requirements are met.

One approach with apparent advantages for NAPLAN is to stage introduction of AES systems. Scoring of the accuracy of student proficiency with the technical skills of writing—such things as punctuation, spelling or grammar—is an obvious place to begin once NAPLAN writing is fully online. Most people believe that a machine can evaluate standard use of writing conventions with acceptable accuracy, as experience with word processing software that does just that shows. This may be more easily accepted than the idea that a computer can 'read' the content and style of an essay. This use of AES for scoring of technical skills may be the most acceptable in the public's eyes, and would offer a significant reduction in the effort required of markers.

Criteria like punctuation and spelling require careful attention from markers, demanding time and increasing cognitive load. Spelling in particular requires markers to count words from classes of word complexity as well as track ratios of correct to incorrect spellings. An AES should be able to mark

these criteria, as well as paragraphing and some or all aspects of sentence structure[14], with a high degree of accuracy and consistency with human markers. In cases where NAPLAN rubrics require markers to count specific instances of a described skill or evaluate a well-defined one, an automated system to replace that effort would likely increase score accuracy while decreasing marking time and cost. It would also allow human markers to concentrate on the compositional aspects of writing, such as ideas, that are not as well supported in the evidence base for automated scoring systems.

The next stage of AES implementation might logically be using humans as the first markers and the AES as a second marker, with a disagreement resolution system where necessary. Putting this into practice would allow for collection of a large data set of AES- and human-scored essays that would be invaluable in completing the research studies needed to support any further uses of the AES. Predictive accuracy could be assessed in fine-grained detail with such a large data set, including any impact on demographic groups of interest or interactions with criteria or writing genre. These data and analyses could be assembled to build an evidence case for uses of an AES in NAPLAN—or to refute one. In order to build or refute the case, acceptance criteria should be developed and standards decided in consultation with stakeholders in the program before any analysis commences. This step should limit the temptation to move the goalposts in either direction during evaluation of the results.

If AES adoption is successful, the data set could also be used towards building some automated feedback systems. One unusual feature of NAPLAN writing assessment is the detailed reporting and performance profiles provided. An AES system could combine the feature extraction in natural language processing (NLP) scoring with the capacity of artificial intelligence (AI) systems to be trained to associate these features with specific types of feedback and recommendations. The training corpus of features and the correct feedback associated with them would need to be developed by expert markers. These markers could locate groups of responses with similar characteristics that would benefit from particular diagnostic feedback or learning recommendations. The AES could be trained using these groupings to provide targeted guidance and critique in NAPLAN reports. Teachers could receive an aggregated report of the feedback given to their students, and perhaps some associated instructional activities that groups of students in their classes might benefit from.

Human markers certainly can provide this type of feedback and guidance but the resource demands to do so are quite high. In his 2005 study, Nichols[15] had markers assign one appropriate annotation as feedback on each essay marked. There were up to 9 possible predetermined annotation choices for each score on a 6-level holistic rubric. Simply choosing an annotation increased marking time by about 40%. This suggests that to provide this sort of feedback in large-scale marking such as NAPLAN, with a substantially more complex rubric, would be time- and cost-prohibitive. However, a well-trained AES might be able to offer enhanced feedback and reporting options beyond those that are feasible with such a large-scale human-marked program.

A further stage might take NAPLAN to an AES as the primary scorer with a human back-marking a proportion of the responses, dealing with the essays rejected by the AES, and resolving any AES-human discrepancies above the threshold size. The human-marked data set could be selected so that it was representative of any demographic or other response groupings of interest. It is useful and necessary to have a human-scored data set for refining AES training. AES require updated training sets, which are human-scored, on a regular basis to assure that they remain tuned to the current standards for writing evaluation.

[14] At a minimum an AES should be able to accurately mark the grammatical and structural components of sentence structure. Marking the 'meaning' aspect of sentence structure with an AES would require careful evaluation.

[15] Nichols, P. (2005). Evaluating the use of annotations when scoring essays. Paper presented at the 35th National Conference on Large-Scale Assessment, San Antonio, TX.

AES will undoubtedly continue to advance in capability and accuracy over time. If NAPLAN takes a staged approach and evaluates each advance in light of program goals, use of AES could reduce cost, burden, and time required for scoring and results reporting while maintaining program standards for accuracy.

5. Research Recommendations

In this section, the individual components above are consolidated into a recommended program of research studies and reviews. A brief summary of the necessary materials and participants is listed, as well as a possible timeline indicating where studies are dependent on results from prior work. Likely risks to data and trends, potential political issues, and advantages to each stage of the described research plan are outlined.

Recommended reviews and studies

1. **Reporting review**: This research should investigate two main facets of NAPLAN writing reporting. One is actual score and report use, including for classroom instruction, by parents and students, by schools and administrators, and education policy impact at the state and commonwealth levels. The other, and possibly more important facet, is desired score use. What is it that stakeholders want the reports to tell them? What would the ideal report design look like for each stakeholder group? What data would it contain? What kinds of displays are most effective in conveying accurate information about the results? What are the primary purposes of assessing writing in NAPLAN and how can those best be served? What kinds of inferences should the data collected in NAPLAN writing support, and where are the appropriate limits of score use? The data for this study could be collected via surveys, focus groups, and interviews with various stakeholders. A design firm should be employed to prototype new reports for consideration as part of this process, as reactions to concrete examples generally are quite informative.
    a. **Goal**: Answers to the questions listed above should direct and inform the design of every aspect of the assessment. Once the goals are clear, the optimal assessment design can be built to serve them – and where there are gaps, it will be known in advance of investment of resources.
    b. **Materials**: Existing NAPLAN writing reports and the *My School* website would be used to investigate the use and interpretation of current reporting. The exploratory phase would require the development and administration of surveys and interview protocols, data collection such as recordings and transcripts from the sessions, as well as qualitative analyses of the responses. Prototype designs would need to be developed based on initial data about use and preferences.
    c. **Dependence/Timeline**: In order to optimise all the work that follows from it, this research should be completed first. This research is not dependent on any other study described here. If significant changes are made in other research studies to NAPLAN writing, parts of this study should be tested against the new designs to assure consistency.
    d. **Risks:** This work is low risk. Focus on the range of NAPLAN reporting materials needed may highlight perceived shortcomings and bring the program under additional scrutiny and may result in unfavourable publicity.
2. **Evaluation of the writing assessment design against the Australian Curriculum: English**. While there have been previous evaluations of the alignment of the Australian Curriculum: English and NAPLAN writing, it may be of added value to conduct an evaluation specifically of the sufficiency of content sampling and the extent to which there is redundancy in the assessment of specific skills. If the overall curriculum or the NAPLAN assessment framework are reviewed, this work is especially important. is reviewed Participants could be drawn across a range of expertise from all stakeholders, including writing instruction experts,

assessment methodologists, current teachers from across year levels, and a range of score users including policy makers, system administrators and parents. The breadth of viewpoints represented in this step is key to maximising buy-in on the outcomes.

   a. **Goal**: The goal of this review would be to determine if the curriculum is appropriately sampled and represented in the assessments, and specifically if the writing curriculum is effectively sampled in the writing assessment as well as other components of the NAPLAN suite of assessments such as reading or conventions of language. The review should be conducted with a view towards the construct span, frequency of assessment of various skills, and adequacy of content representation.
   b. **Materials**: The data and materials needed for this work largely are readily available. Training materials for the evaluators would need to be created as well as forms for data capture. These tools maximise the benefit of the review by assuring that responses and feedback are captured in a consistent way and that no aspects of the desired review is neglected.
   c. **Dependence/Timeline**: This work is not directly dependent on any other and likely should be completed first as it may provide direction and focus to later efforts. The work of recruiting stakeholders for participation and assembly or creation of materials could commence immediately.
   d. **Risks**: This work is low risk. Access to the range of NAPLAN materials needed may bring the program under more focused scrutiny and may result in unfavourable publicity. However, currently NAPLAN faces intense public and policy interest and criticism and this work is not likely to dramatically exacerbate that.

3. **Audit of rubrics, criteria and exemplars**: As indicated in Section 2 of this report, there are a number of aspects of the current NAPLAN writing rubrics, criteria and exemplars that would benefit from a careful review. The review should incorporate as many informed viewpoints as is feasible to maximise the impact of the work. Suggested for inclusion would be markers and marking leadership, writing instruction experts, methodologists with expertise in human marking and rubrics, and current teachers from across year levels.
   a. **Goal**: The goal of the review would be to revise the current writing assessment materials into a more focused and coherent set. This review should consider the number of criteria assessed; the overlap between criteria; other sources of skills assessment data; and optimisation of the number of criteria and score levels in each to assure they are distinct. For each criterion, the key skills assessed and the evidence used to support level assignment should be made explicit. The language and usage within and across criteria should be made consistent. Sample scripts should be reviewed and supplemented or decreased as best to align with the revisions recommended. Where vague quantifiers are used, sample scripts should be selected that exemplify the distinctions intended. If explicit quantification is included, the rationale for requiring these counts should be made explicit. Where score categories are used infrequently, consideration should be given to removal or combining adjacent categories. As part of this review, consideration of the efficacy of separating the rubrics so that different year levels utilise different rubrics as well as the pros and cons of combining the current rubrics into one NAPLAN writing rubric used across text types should be included.
   b. **Materials**: The data and materials needed for this work largely are readily available. Participants should meet regularly to discuss progress and assure a consolidated set of recommendations at the conclusion of the work.

c. **Dependence/Timeline**: This work is not directly dependent on any other but likely should be completed in conjunction with the curricular alignment study in 2, as the combined results will provide direction and focus to later efforts. The work of recruiting stakeholders for participation and assembly or creation of materials could commence immediately. Note that if changes are made to the rubric, criteria, and/or exemplars, this will have a direct impact on current and new marker training components and marking quality assurance measures.
d. **Risks**: This work is low risk. Implementation of recommended changes (considered below) may increase the risk of public perception issues.

4. **Marking design review:** There are a number of design facets of NAPLAN writing marking recommended for consideration and change outlined in Section 3 of this report. A comprehensive review should be undertaken to consider which of the possible modifications to NAPLAN marking would be considered most desirable by stakeholders, be feasible to implement given political and budgetary constraints, and maximise the return of value for investment of resources. The reviewers should include key stakeholders such as markers and marking leadership as well as policymakers at the state and commonwealth level. This work should be led by experts in marking and in software system design so that any recommended alterations are both supported by research and practice in the field and viable for implementation in training and marking systems.
    a. **Goal**: The review should target such aspects as: elimination of face-to-face marker training variability by moving marker training online; creation and implementation of bias reduction training as part of marker training; reviewing the within-state marking design in light of the bias that may be induced by it (this bias could be evaluated in a separate study if desired); creation and standardisation of marker fatigue rules for the program; and a consideration of piecework versus hourly pay rates and its interaction with the inclusion of quality assurance measures. Quality assurance structures to be considered may include: a review to decide which quality assurance measures are needed and useful for NAPLAN specifically; selection of the necessary exemplars and quality control cases, including activities such as expert selection and rationale development; and creation of software design specifications for the marking system. Once the QA systems have been delineated, program changes to be considered include: standardisation of quality assurance procedures across state marking centres through procedures such as requiring all markers to re-certify in every cycle regardless of prior experience; procedures for regular marker skills assessment and quality-assurance case use across all marking centres; and consistent standards and consequences for markers who are unsuccessful in QA measures.
    b. **Materials**: Much of the data and materials needed for this work largely are currently available.
    c. **Dependence/Timeline**: The curricular alignment and rubric audit should be completed prior to this work commencing so that the benefits of that work are incorporated into this review.
    d. **Risks**: This work is low risk. Implementation of recommended changes (considered below) may increase the risk of public perception issues.
5. **Impact of re-weighting criteria on scores and reporting**: This study could be narrowly focused on decisions about how NAPLAN criterion weights should be distributed. Expert opinion could be invited to decide on a small set of theoretically-based models for reweighting the existing writing scores *post hoc*. A possible model might include classification of criteria

into groups (*e.g.* conventions, authorial, and structural) and decisions on the relative value each contributes to the overall construct.

   a. **Goal**: The target of interest is the magnitude of the impact such a re-weighting would have on the formative information provided to students and teachers, classification into performance bands, performance trends over time, and relative ranking of schools.
   b. **Materials**: The classification exercise and analysis work both utilise existing and available data and materials.
   c. **Dependence/Timeline**: This work is not directly dependent on any other listed here.
   d. **Risks**: This work is low risk. Operational implementation of any changes may increase the risk of public perception issues.

6. **Assessment design review**: There are a number of aspects of the current NAPLAN assessment design that intersect with the marking design. These include the use of a single task for assessment; possible inclusion of additional text types; a redesign to incorporate multiple texts and tasks; and varying specific texts and task within student year level. Changes to some or all of these aspects would suggest the need to break scale and re-start the measurement trend. This could be done at the move to fully online assessment. Although it is out of scope for this paper, there also are issues in analysis and equating that could be addressed effectively if the reporting scale is re-set.
   a. **Goal**: The target of the assessment design review could be relatively narrow or extremely broad. More narrowly, the focus would be on construct sampling in writing, specifically whether the administration of a single text and type should be re-considered. If the initial review concludes that the current assessment is not adequate, then a second phase might explore the ramifications of that decision in terms of a more complete redesign of the assessment.
   b. **Materials**: The necessary materials needed for this work are currently available.
   c. **Dependence/Timeline**: While this work is not directly dependent on any prior studies, it is probable that the results from Studies 1 through 5 above could influence the evaluation of the assessment. It is thus recommended that the results of these studies be available for this one.
   d. **Risks**: The review is low-risk by itself. Depending on the breadth of the revisions undertaken, this work incurs high risk of public and political repercussions. Significant changes to such a visible assessment as NAPLAN will inevitably be unpopular in some quarters. The media may cast this work not as an improvement but as rectification of a mistake and potentially damage the credibility of the NAPLAN program more broadly. Loss of trend measurement would likely be unpopular with score users at all levels.

7. **Revised assessment and marking validation and implementation (series of studies):** The work above will require a series of studies to build the evidence argument for valid use of the revised assessment. The initiating work described in Studies 1-6 will define the parameters of this work. Assuming that some set of modifications or re-designs have been selected to investigation for operational use, a design for a single impact study of the set of changes or a series to studies to disentangle the impact of individual adjustments may be necessary.
   a. **Goal**: Detailed targets and outcomes for this phase of work are dependent on the prior efforts and the changes selected for trial and possible implementation.
   b. **Materials**: Rubric revisions could be evaluated using existing essay responses if the revisions are limited in scope and no modification of the essay content or directions

are contemplated. Marking design changes also could be evaluated using existing essay responses if the revisions are limited in scope and no modification of the essay content or directions are contemplated. If the marking design requires new or modified software systems, these would have to be specified and built, as well as trialled for operational fitness. Potentially, new items will be developed to accommodate aspects such as modified timings, varying text types, and/or multiple text per student that will require trialling and revision if selected.

c. **Dependence/Timeline**: If a significant revision to the assessment design is contemplated, that should be completed before the other aspects of the work are considered for validation. There is no point in revising the current rubrics if the items will no longer be aligned to that type of scoring.

d. **Risks**: The studies are not inherently high risk. The commitment of resources necessary to complete them successfully is a risk that can be mitigated by careful planning of each stage and thorough review of each set of outcomes before the next stage commences. As with Study 6, this work incurs high risk of public and political repercussions and for similar reasons.

8. **Evaluation of modifications made on dependence:** In Section 3 of this report there is a fairly comprehensive set of evidence of substantial dependence in the data of NAPLAN writing. This finding is not new and has been reporting in a variety of other research as well. As noted, there is not an expectation that the dependence can be removed analytically; indeed, analytic approaches implemented herein had little or no effect on the degree of data dependence observed. The most probable path to a reduction in this aspect of the data runs thorough many of the studies described above, especially 3, 4, and 6. In investigation of the impact of these changes on the degree of observed over-consistency in marking should be undertaken as part of the validity evidence argument construction for the assessment as a whole, described briefly above in Study 7.

    a. **Goal**: This investigation would evaluate any reduction in the degree of data dependence observed in the NAPLAN data. There are numerous approaches to do this, so a thoughtful selection of specific approaches should be made.

    b. **Materials**: The data necessary to investigate this should be collected as part of the work of Study 7.

    c. **Dependence/Timeline**: Completion of students 3, 4, 6 and parts of 7 are required prior to commencing this work.

    d. **Risks**: There is a risk that, should this study find that the dependency is the same or has been increased by the changes, the finding would undermine the work done towards improvement of the writing assessment program writ large. It is virtually impossible that all dependence will be removed from any performance assessment regardless of design. Nonetheless, quantifying the degree to which dependence remains after program changes may be seen as a condemnation of the assessment.

9. **Automated essay scoring**: As described in Section 4.5 of this report, NAPLAN is well positioned to consider a staged implementation of automated essay scoring systems. It is important to note that there exists a body of high-quality research into implementation of AES in NAPLAN writing specifically. Such research consistently provides more than adequate technical support for such an implementation. Given repeated findings of statistical accuracy and adequacy for multiple AES systems, it is somewhat surprising that this approach has not been implemented already. Since technical factors do not appear to be the limitation, it seems probable that human factors such as negative public reactions may be the barrier.

The multiple criteria used in the NAPLAN writing rubric lend themselves well to partitioning the marking task into components that AES systems may be seen to do well, such as grammar, spelling, vocabulary, and punctuation, and authorial criteria that may remain with humans to mark. This implementation would reduce the load on the human markers, allowing them to concentrate their attention on the more-subjective aspects of writing. It may also alleviate perceptual concerns about 'robots marking writing'. One aspect of such an implementation that should evaluated in this study is the impact on human marking of this AES reduction. The design should incorporate selected scripts that allow for examination of whether the human markers are able to ignore the mechanical aspects of writing that the AES is marking and concentrate on their assigned criteria, or whether the language conventions aspects are intrusive and influential even when not part of the human-scored rubric. Once AES systems are implemented and visibly functioning well as part of NAPLAN marking, the transition to using such systems more broadly may encounter less resistance.

a. **Goal**: The goal of this research is to evaluate the viability of a split marking process, where the AES initially is limited to the language conventions criteria. This is expected to provide a reduction of the current human marker load as well as savings of resources while maintaining data quality, reporting standards, and accurate formative feedback.
b. **Materials**: AES systems require a representative set of key-entered essays marked by expert human markers in order to train the models and produce accurate results. If changes are made in the above series of work, a new set of these data would be required before AES implementation could occur. The accuracy of the AES systems has been established previously using responses and rubric in the current NAPLAN data set; the human ability to effectively mark the reduced set of criteria and ignore the other is the focus of investigation.
c. **Dependence/Timeline**: The dependencies for this study differ based on whether the introduction is desired before or after the revisions and reviews recommended above. If before, the study could be designed and operationalised within the next 6-12 months, as all necessary materials were assembled for the previous AES trials. Materials would need to be developed to train the human markers on the reduced rubric and to emphasise not incorporating language conventions as evidence. AN advantage of access to the previous AES and human-marked data set using the full NAPLAN rubric in both cases is allowing evaluation of the reduced rubric on human marking (it is not expected that the reduction will alter the quality of the AES as previously established). If AES introduction is reserved until after the other revisions above, then a new evaluation study of AES efficacy on the new assessment would be required.
d. **Risks**: As noted above, the risks to AES introduction in NAPLAN are primarily human and perceptual, not technical. While it is reasonable to infer that AES introduction in a targeted and more-narrow way might be perceived differently than full AES marking, it is always possible that any AES system will result in negative publicity and political pressures.

The described program of work would encompass several years of effort. It is likely to result in a substantially different program for assessing writing in NAPLAN than the one that currently exists. The studies themselves would help guide those changes so that they result in a program that has a clear purpose; informative reporting; validated uses for scores; and optimised designs for the

assessment, the marking rubric, and the work of operational marking. Future advances in all aspects of the assessment of complex performances such as writing will undoubtedly alter the course of research described in this section, just as current knowledge has influenced the recommendations made here. NAPLAN is an influential large-scale assessment program that incorporates uncommon expectations like formative information as well as more usual uses such as trend measurement and influence on education policy. Maintaining that position is a balance that would be well-served by a wide-ranging, carefully planned program of research and evaluation that keeps the assessments at the forefront of practice.

6. Appendix A: Writing Marking Rubric Examples

ISA NARRATIVE/REFLECTIVE SPELLING

SCORE	DESCRIPTION
09	The spelling of **a wide-ranging, mature** vocabulary is virtually error free. Vocabulary is appropriate to the context.
08	The spelling of **a carefully selected vocabulary** is correct most of the time. • has a few errors within a range of conventional vocabulary
07	The spelling of **a student vocabulary** is well-controlled. • spells correctly a number of words with more difficult patterns • usually chooses correct spelling of homophones
06	The spelling of **a wider range of vocabulary** commonly used by school students is mostly correct, though first draft writing shows some uncertainty or inconsistencies. • spells correctly some difficult words • spells correctly some contractions or less common words
05	Most spelling, **within a limited student vocabulary**, is usually correct. • May spell correctly a number of longer words, for example, compound words • May spell correctly some words containing silent letters • May successfully use some spelling rules
04	The spelling of a **simple vocabulary** is mostly correct. Spelling of a wider selection of vocabulary may be inconsistent.
03	The spelling **supports**, **rather than interferes with**, the reading of the texts. Shows awareness of phonetic and visual patterns.

SCORE	DESCRIPTION
02	The spelling **makes writing difficult to read**, though some simple words may be spelt correctly.
01	Very few identifiable words. **May require interpretation** of intended letters and words.
IE	**Insufficient evidence to judge**. Spellcheck used (must have evidence from information provided)
Missing	No response

Notes for markers:

- **when scoring, read the descriptions from low to high**
- **Any description of a *positive* feature of writing at one level is assumed (as a minimum) to exist also in the levels above it**

LEVEL	DESCRIPTION	POINTERS
10	The writing is **mature and fluent**. Well-constructed sentences are polished in grammar, syntax and punctuation. A wide range of vocabulary is used effectively. The writing displays confidence and control through a well-established voice and attractive style. Dialogue, if used, sounds authentic and is properly punctuated.	• fluent, mature writing • all aspects of language meet demands of the narrative/reflective genre • strong individual voice and flair • wide and effective use of vocabulary • vocabulary is sophisticated but not pretentious
09	The writing demonstrates **competent use of standard English**. A coherent structure is used effectively. Expressive and imaginative vocabulary and phrasing enhance the writer's ideas or create mood and atmosphere. The writer's individual voice and style show a growing maturity.	• competent use of standard English; fluent and smooth • coherent structure • noticeable voice and style (though may be a little uneven) • effective and imaginative choice of vocabulary and phrasing
08	The writing shows **general control of standard English**, with correct grammar and punctuation. The overall structure and organisation are appropriate to the narrative/reflective genre. Vocabulary selection is competent. The writer's voice is evident.	• general control of standard English; fluent • writer's voice is evident • correct grammar and punctuation • competent vocabulary selection

LEVEL	DESCRIPTION	POINTERS
07	The writing shows **control of grammar and punctuation in a variety of complex sentences**. Organisation of text is competent; may include paragraphing or other appropriate features. Vocabulary is precise or selected for effect, although perhaps not sophisticated or extensive in range. Emerging voice can be recognised.	• control of variety of sentence structures; degree of fluency • competent text structure • precise use of selected vocabulary • emerging voice
06	The writing **generally shows control of grammar and punctuation in complex sentences**. Organisation of text is appropriate. Standard or conventional vocabulary is used appropriately. There may be an indication of emerging voice.	• general control of grammar and punctuation • appropriate text organisation • conscious vocabulary choice • indication of emerging voice
05	The writing is **generally fluent and smooth**, though there may be occasional lapses in grammar and syntax. It includes a variety of simple, compound and complex sentences, using a range of conjunctions and other linking devices and suitable punctuation. Paragraph divisions or other appropriate organisational features may be used. Vocabulary suits the content and text type.	• generally fluent • variety of sentence forms with appropriate punctuation • vocabulary suits content • possible lapses in grammar and syntax

LEVEL	DESCRIPTION	POINTERS
04	The writer is **becoming fluent**. The writer uses **compound and complex sentences** in which clauses are joined by linking words, for example, when, after and because. Varied sentence beginnings are used. Evidence of conscious vocabulary selection. May include a range of common punctuation.	• becoming fluent • attempt to select vocabulary for effect • attempt to vary sentence forms/beginnings • may attempt direct speech • may begin to control punctuation
03	The writing shows **control of simple sentence structure**. The sentence shape is clear. A simple vocabulary is used. Sentences are often linked by 'and', 'but', 'so', 'then'.	• control of simple sentences • basic vocabulary • may be 'run on' sentences
02	Writing uses **basic conventions**, attempts to use simple sentence forms and simple vocabulary.	• attempts simple vocabulary and sentence forms
01	Very **basic language** used; very little meaning can be gleaned	• attempts to follow the most basic conventions • some recognisable letters or words • writes from left to right
IE	**Insufficient evidence to judge** For example clusters of letters, pictures and invented forms of writing.	• includes drawings, foreign language, copying of prompt

LEVEL	DESCRIPTION	POINTERS
Missing	No response	

Notes for markers:

- **when scoring, read the descriptions from low to high**
- **Any description of a *positive* feature of writing at one level is assumed (as a minimum) to exist also in the levels above it**

ISA NARRATIVE/REFLECTIVE CONTENT

LEVEL	DESCRIPTION	POINTERS
10	The **writing is sustained and presents a complex and mature reflective viewpoint or approach**. If a narrative, the writer may adopt and sustain a convincing persona as author or participant in the action. The skilfully constructed piece displays originality and is supported by carefully selected detail. If a narrative, characterisation shows emotional or psychological complexity. The writing evokes a strong response in the reader.	• sustained narrative / reflection with complexity of purpose, sophisticated approach or subject matter • thought-provoking reflection on attitudes, values or issues • writer adopts convincing persona • emotional or psychological depth to characters • evokes strong response in reader • likely to go beyond stereotypes
09	**A carefully constructed piece that may reflect on values and offer insights**. There may be some reflection underpinning or implicit in the piece. If a narrative, characters are credible, with the reader given insight into their lives. Relationships between characters are convincing.	• sustained and unified /reflection with a well-constructed conclusion • may be more than a linear construction, for example, more than one complication • empathetic response to characters • reflects on attitudes and values • reader's interest strongly caught
08	The writing **is a developed piece**. The **overall structure is appropriate with a clear direction.** **(**may be unfinished**)** If a narrative, characterisation is credible, for example, through the presentation of motive underpinning action or emotional response to the situation. If a reflection, the writing conveys genuine engagement with the task and the reader.	• developed and integrated • credible character development/reflection • sure as a narrator • attention to time order • engages reader

LEVEL	DESCRIPTION	POINTERS
07	The writing **is a well-constructed piece within a sound structure.** A deliberate intention of engaging the audience is evident. Ideas and events are appropriately linked. If a narrative, the characterisation is credible, with characters clearly individualised. If a reflection, the writer's point of view is clear.	• sound structure • sense of voice • generally sound characterisation (narrative) • clear point of view (reflection)
06	The narrative / reflection has a **clear sense of purpose** The writing contains ideas, details and events chosen to enhance the piece of writing. Characters are distinguished either explicitly through description or implicitly through action and speech	• focus maintained • characters clearly defined • a degree of detail • sense of audience
05	The writing shows an **understanding of the narrative/reflective genre**. Many ideas contribute to the storyline or reflective thread, but the piece may fall away or lack resolution. If a narrative, a sense of the character or characters begins to emerge through description or through actions and speech. There is a conscious consideration of audience, for example, an attempt at mystery, suspense, adventure, fantasy or other genres	• developing skill in plot construction or reflection (may fall away) • characters emerging • some conscious consideration of audience

LEVEL	DESCRIPTION	POINTERS
04	**A notion of the story /reflection as a whole** is evident in an attempt to shape the piece. The writing contains key events, main characters (if a narrative), and a setting. Characters, while introduced in the story, are not well defined.	• distinguishable storyline / reflective thread • minimal character definition • emerging sense of audience • story setting is clear; logical sequence of events
03	**Simple storyline** or reflective thread. May be episodic. Characters, if present, may be named but not individualised. There is little evidence of selection and control of the content to achieve a specific purpose.	• limited attempt at audience impact • characters named only or undeveloped ideas • episodic
02	**Awareness of task**. Skeletal story. No clear development or shape.	• some understanding of task • very basic elements of story/reflection; may be brief, but lacks coherence • may contain suggestion of character
01	**Minimal response.** May have written a short sentence.	• indication of an attempt to write something, but communicates little to reader • little to assess
IE	**Insufficient evidence to allow meaningful judgement** Evidence that the student was present;	• includes drawings, foreign language, copying of prompt, plan only
Missing	No response	• no evidence of the student being present

Notes for markers:

- **when scoring, read the descriptions from low to high**
- **Any description of a *positive* feature of writing at one level is assumed (as a minimum) to exist also in the levels above it**

MTEG Items[16]

MED Afghanistan, WRITING

Scenes We See (2)

Write two sentences to describe each picture.

1. .....
.....
.....

2. .....
.....
.....

1. .....
.....
.....

2. .....
.....
.....

Title W0029S01P

 15

[16] Items have been used in Afghanistan and are intended for use at Grade 6. MTEG materials provided here are not secure.

MTEG Task List

Task ID	Task name	booklet/page	booklet/page	time	cluster	Task name
W0001S01P	Making Life Better	B1/34	B2/33	15	W1	**Making Life Better**
W0002S01P	Amina	B2/30	B3/36	5	W2	**Amina**
W0004S01P	Mice in the Kitchen	B1/36	B6/36	10	W6	**Mice: Main picture**
W0004S02P	Mice in the Kitchen	B1/37	B6/37		W6	**Mice: 3 sentences together**
W0004S03P	Mice in the Kitchen	B1/37	B6/37		W6	
W0004S04P	Mice in the Kitchen	B1/37	B6/37		W6	
W0008S01P	How to Grow Beans	B3/34	B4/34	7.5	W3	**How to grow beans**
W0010S01P	Inviting Uncle	B5/33	B6/39	10	W5	**Inviting uncle**
W0012S01P	Visiting Cousin	B3/35	B4/35	7.5	W3	**Visiting cousin**
W0015S01P	New Student	B2/32	B3/37	5	W2	**New Student**
W0021S01P	The Bird and the Box	B4/32	B5/34	15	W4	**The bird and the box**
W0026S01P	Objects we see	B1/35	B6/34	5	W6	**Objects we see**
W0028S01P	Scenes we see 1	B5/32	B6/38	5	W5	**Scenes we see 1**
W0029S01P	Scenes we see 2	B2/31	B3/37	5	W2	**Scenes we see 2**

MTEG Marking Guide by Task

Task ID	Task name	time	criterion	score	description
W0001S01P	**Making Life Better**	15	W0001T02P quality of ideas / argument	0	evidence of a response, but no relevant information is included
				1	repeats or paraphrases prompt/task; no new ideas
				2	minor elaboration
				3	some elaboration; may include some ideas somewhat irrelevant to the task
				4	well elaborated and clearly relevant to the task, answering the question asked. There is some complexity or broadness in the thinking.
			W0001T04P syntax & grammar	0	isolated words or sentence fragments OR copied
				1	some sentences are incomplete OR sentences contain many errors
				2	sentences are very simple and repetitive but generally correctly formed OR are more complex but with errors
				3	sentences are varied in structure and correctly formed
			W0001T01P coherence/ cohesion	0	evidence of a response, but unintelligible OR copied
				1	ideas are disjointed (not related to each other), so the text is not easy to follow
				2	ideas generally follow a logical sequence but are not adequately linked with connecting words; meaning is relatively easy to follow
				3	ideas are well related and easy to follow throughout
			W0001T03P spelling	0	can only spell words given in prompt
				1	can spell basic words
				2	shows ability to spell beyond basic words
W0002S01P	**Amina**	5	W0002T02P vocabulary	0	no relevant verb or adjective
				1	verb <u>or</u> adjective/description of feeling: skipping/jumping/ playing/ smiling; happy
				2	verb <u>and</u> adjective/description of feeling

Task ID	Task name	time	criterion	score	description
			W0002T01P grammar	0	evidence of a response but no grammatically correct sentences
				1	1 grammatically correct sentence (no errors)
				2	2 grammatically correct sentences
W0004S01P	**Mice: Main picture**	10	W0004T01Poverall account of the picture	0	evidence of a response, but no relevant information is included
				1	focuses on isolated features or elements Or is very general / superficial
				2	gives a good sense of the whole picture and includes some detail
			W0004T02P sentence structure	0	isolated words or sentence fragments
				1	at least one simple correct sentence; sentences may be repetitive in structure
				2	Generally there are at least 2 grammatically correct sentences. One or more of the sentences are complex or sentences are varied in form
			W0004T03P vocabulary the student uses	0	no relevant content words
				1	(limited range): basic vocabulary, repetitive, inadequate to describe the content the student presents
				2	adequate to describe the content the student presents
				3	good range of vocabulary gives good sense of detail of the content the student presents: good range of verbs and adjectives and nouns
W0004S02P	**Mice: 3 sentences together**		W0004T04P grammar and word choice	0	no appropriate and correctly formed sentences
W0004S03P				1	1 relevant and correctly formed sentence (eg singular/plural agreement).
W0004S04P				2	2 relevant and correctly formed sentences. Sentences must be different
				3	3 relevant and correctly formed sentences. Sentences must be different
			W0004T05P verb formation	0	does not use present tense and person/subject correctly

Task ID	Task name	time	criterion	score	description
				1	uses present tense and person/subject correctly in all sentences that are written (may be 1, 2 or 3 sentences). May use habitual or present tense.
W0008S01P	**How to grow beans**	7.5	W0008T01Pinstructional language	0	some writing, but no evidence of instructional language
				1	some use of imperative or other instructional language
				2	consistent use of imperatives or other instructional language
			W0008T02P relevant information	0	insufficient ideas to get meaning across
				1	process is not clearly presented
				2	process is clearly presented
			W0008T03P spelling	0	can only spell words given in prompt
				1	can spell basic words
				2	shows ability to spell beyond basic words
W0010S01P	**Inviting uncle**	10	W0010T02P persuasive	0	writing but no sense of persuasion
				1	some attempt at persuasion using relevant reasons, but not convincing
				2	letter is convincing
			W0010T01P correct sentences	0	isolated words or sentence fragments OR copied
				1	some sentences are incomplete OR sentences contain many errors
				2	sentences are very simple and repetitive but generally correctly formed OR are more complex but with errors
				3	sentences are varied in structure and correctly formed
			W0010T04P spelling	0	can only spell words given in prompt
				1	can spell basic words
				2	shows ability to spell beyond basic words
W0012S01P	**Visiting cousin**	7.5	W0012T02P ideas (relevance)	0	some writing but nothing seems relevant; not a message
				1	some relevant ideas but not enough for an adequate note / message
				2	note is complete in terms of ideas and message
			W0012T03P vocabulary	0	little control of relevant vocabulary
				1	vocabulary used shows limited ability to convey a message

Task ID	Task name	time	criterion	score	description
				2	vocabulary is adequate to convey detail of message
			W0012T01P handwriting	0	few letters are well formed
				1	legible, most letters well formed
				2	good control of letter formation throughout
W0015S01P	**New Student**	5	W0015T03Prelevant ideas	0	no ideas are relevant or interpretable
				1	1 or 2 interpretable ideas relating to the new student
				2	3 interpretable ideas relating to the new student
			W0015T01P correct question form	0	no correct question forms
				1	1 or 2 correctly formed questions or other appropriate correctly formed sentences/responses (eg 'I would ask about …')
				2	3 correctly formed questions or other appropriate response that includes 3 questions.
			W0015T02P handwriting	0	few letters are well formed
				1	legible, most letters well formed
				2	good control of letter formation throughout
			W0015T04P punctuation	0	no evidence of ability to use question marks where they are needed
				1	question marks are used where needed OR if they are not needed, other appropriate punctuation is used (full stops)
W0021S01P	**The bird and the box**	15	W0021T05Pstory elements	0	evidence of a response, but no relevant information is included. No introduction or conclusion.
				1	has an introduction (scene setting) or an ending (conclusion)
				2	has an introduction (scene setting) and an ending
			W0021T02P narrative sequence	0	evidence of a response but no relevant information is included
				1	ideas are present but not a narrative
				2	ideas are linked into a narrative
			W0021T01P elaboration of ideas	0	evidence of a response, but no relevant information is included
				1	fragments: few ideas or no complete ideas

Task ID	Task name	time	criterion	score	description
				2	limited writing related to the picture
				3	simple writing related to the picture; limited detail
				4	detailed writing with many relevant ideas
			W0021T03P punctuation	0	no evidence of ability to use punctuation (no commas or full stops correctly used)
				1	some correct use but some problems with punctuation
				2	correct use of punctuation
			W0021T04P sentence structure and complexity	0	isolated words or sentence fragments OR copied
				1	some sentences are incomplete OR sentences contain many errors
				2	sentences are very simple and repetitive but generally correctly formed OR are more complex but with errors
				3	sentences are varied in structure and correctly formed
W0026S01P	**Objects we see**	5	W0026T02Pvocab	0	wrong word or not recognisable
				1	recognisably correct: foot
			W0026T01P spelling	0	incorrect spelling
				1	correct spelling
			W0026T08P vocab	0	wrong word or not recognisable
				1	recognisably correct: tree / bush
			W0026T07P spelling	0	incorrect spelling
				1	correct spelling
			W0026T06P vocab	0	wrong word or not recognisable
				1	recognisably correct: glass / cup / vase
			W0026T05P spelling	0	incorrect spelling
				1	correct spelling
			W0026T04P vocab	0	wrong word or not recognisable
				1	recognisably correct: fire / flame / heat / coal / cooking
			W0026T03P spelling	0	incorrect spelling
				1	correct spelling
W0028S01P	**Scenes we see 1**	5	W0028T03P vocabulary	0	fewer than 3 relevant content words (verbs, nouns, adjectives)
	donkey			1	at least 3 different relevant content words (nouns, verbs, adjectives)

Task ID	Task name	time	criterion	score	description
			W0028T01P syntax / sentence structure	0	isolated words or sentence fragments only
				1	some errors but comprehensible OR one simple sentence correctly formed
				2	2 simple sentences, correctly formed, or one complex / compound sentence correctly formed
	football		W0028T04P vocabulary	0	fewer than 3 relevant content words (verbs, nouns, adjectives)
				1	at least 3 different relevant content words (nouns, verbs, adjectives)
			W0028T02P syntax / sentence structure	0	isolated words or sentence fragments only
				1	some errors but comprehensible OR one simple sentence correctly formed
				2	2 simple sentences, correctly formed, or one complex / compound sentence correctly formed
W0029S01P	**Scenes we see 2**	5	W0029T03P vocabulary	0	fewer than 3 relevant content words (verbs, nouns, adjectives)
	bird			1	at least 3 different relevant content words (nouns, verbs, adjectives)
			W0029T01P syntax / sentence structure	0	isolated words or sentence fragments only
				1	some errors but comprehensible OR one simple sentence correctly formed
				2	2 simple sentences, correctly formed, or one complex / compound sentence correctly formed
	cars		W0029T04P vocabulary	0	fewer than 3 relevant content words (verbs, nouns, adjectives)
				1	at least 3 different relevant content words (nouns, verbs, adjectives)
			W0029T02P syntax / sentence structure	0	isolated words or sentence fragments only
				1	some errors but comprehensible OR one simple sentence correctly formed
				2	2 simple sentences, correctly formed, or one complex / compound sentence correctly formed