acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

NAPLAN ONLINE AUTOMATED SCORING RESEARCH PROGRAM:

RESEARCH REPORT

Dr Goran Lazendic, Julie-Anne Justus and Dr Stanley Rabinowitz

Executive summary

The purpose of this paper is to present details of the National Assessment Program – Literacy and Numeracy (NAPLAN) Online automated scoring research program and its key outcomes. The research program, developed and led by ACARA, was conducted in collaboration with Pacific Metrics, the provider of the Constructed-Response Automated Scoring Engine (CRASE®).

The research is designed to collect and evaluate empirical evidence on the feasibility and validity of automated scoring for NAPLAN writing assessments, based on a range of studies and analyses. The studies and analyses presented in this report used a broad, nationally stratified sample of over 11,000 essays across eight persuasive and four narrative writing prompts. Current industry and research standards were used to evaluate the performance of CRASE® to determine whether its scoring is consistent with that provided by human markers. The research also focused on some of the key aspects of NAPLAN writing assessments and whether its underlying measurement construct, including the design and implementation of the NAPLAN marking rubric, is conducive to automated scoring. Finally, the research investigated some of the practical issues of potential implementation of automated scoring.

Research findings demonstrated that the modern automated scoring system tested, when marking NAPLAN writing, provided the same level of reliability and consistency as that found between two independent sets of human markers. Further results showed that CRASE® was resilient to attempts to manipulate marking and that the latent structure of automated scores was the same as that of the human markers.

Acknowledgements

Martina Bovell and Ross Bindon (ACARA), and Dr Susan Lottridge and Dr Scott Wood (Pacific Metrics) contributed their expertise, initiative and effort to the successful implementation and execution of the NAPLAN Online automated scoring research program. Most statistical analyses reported in this report were completed by Pacific Metric in collaboration with ACARA. ACARA also thanks members of its Measurement Advisory Group (MAG) for their valuable feedback on this report.

Introduction

The National Assessment Program – Literacy and Numeracy (NAPLAN) is an annual assessment of reading, writing, convention of language and numeracy for all students in Years 3, 5, 7 and 9. The assessments are undertaken nationwide in May, after the end of the first three months of a school year. The purpose of the administration of NAPLAN tests this early in a school year is to provide data as early as possible so that these data can be used to improve student learning outcomes. One of the criticisms of NAPLAN is that the results of the current paper-based tests are released too late in a school year to be effectively used to inform and support teaching and learning.

In addition to providing better and more precise assessment, moving NAPLAN online will significantly shorten the time needed to provide results to schools and students by removing the administrative and logistic burden of processing millions of paper tests and conducting longitudinal equating of NAPLAN assessment scales prior to administering the tests in May. During the agreed transition period, schools that switch to NAPLAN Online will receive a first set of their test results three weeks after the online test window is closed. The turnaround of NAPLAN results could be reduced to several days. The new online and adaptive NAPLAN tests in reading, conventions of language and numeracy will be computer-scored as a student progresses through the tests, allowing drastically reduced time to provide results and feedback on the NAPLAN online tests.

Marking of NAPLAN writing tests is currently a large logistical operation that takes several months and requires significant resources to set up marking centres, employ, train, manage and monitor a large number of markers. Marking of approximately 900,000 NAPLAN writing scripts takes significant time. It is to address this delay that automated scoring of NAPLAN writing is being investigated.

Automated scoring of writing relies on computer learning and modelling methods to develop automated scoring models that emulate outcomes of human markers' scoring. Current literature provides evidence for the feasibility of implementation of automated scoring of large-scale writing assessments (Lochbaum et al., 2015; McGraw-Hill Education CTB, 2014; Shermis, 2014) as well as opposing views regarding the concept and validity of automated scoring (Haswell & Wilson, 2013; Perlman, 2013, 2014). Unfortunately, a large proportion of this research debate occurs in a context in which the interaction of automated scoring with the underlying measurement concept of a writing assessment is not always well-defined. Scoring of writing assessments is inherently linked to its underlying measurement construct and thus this relationship must be clearly defined in all its aspects to provide a fair evaluation of validity evidence for automated scoring in different writing assessments – from the design of the writing prompts to design and implementation of marking rubrics.

NAPLAN writing tests are designed to assess progression of knowledge and skills that make students effective written language communicators. The knowledge assessed includes knowing of the purpose of a text, its audience, different text types, as well as a command of lexical and syntactical aspects of English, set in the Australian Curriculum. The underlying construct of NAPLAN writing tests is thus independent from a year level and it explicitly recognises the progressive development of written language. Such understanding and definition of the writing concept are recognised in the construction of NAPLAN prompts that include two year levels – Year 3 and 5, and Years 7 and 9 respectively – and in the design of NAPLAN marking rubrics that apply to all prompts irrespective of year level. The marking rubrics recognise that different text types require different combinations of the elements of written language and thus separate marking rubrics are developed for persuasive and narrative writing genres. Finally, the purpose of NAPLAN marking rubrics is to communicate which elements of students' writing require further attention and to understand the students' literacy learning outcomes and needs. To that end, NAPLAN marking rubrics have 10 criteria:

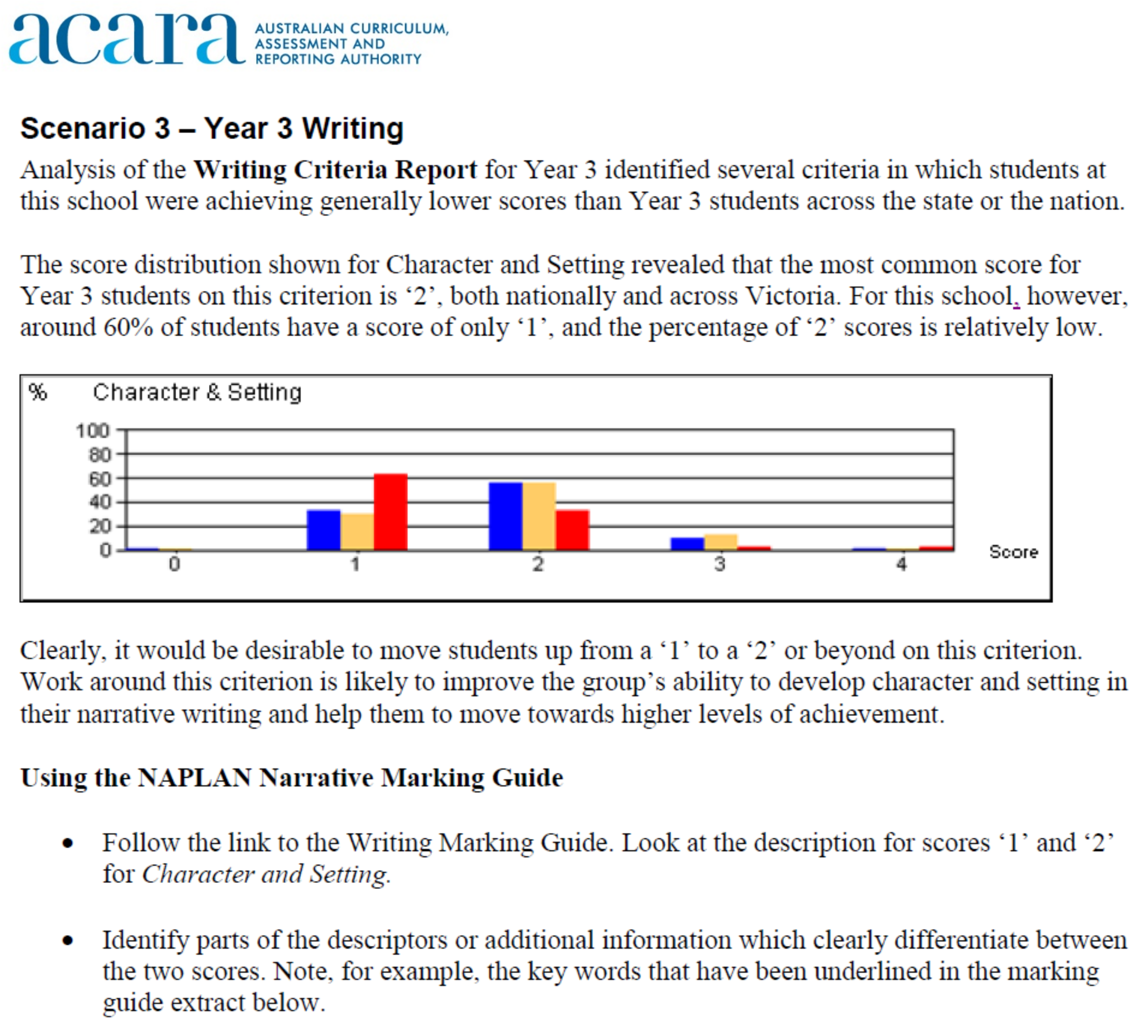
1. audience
2. text structure
3. ideas
4. character and setting (narrative), persuasive devices (persuasive)
5. vocabulary
6. cohesion
7. paragraphing
8. sentence structure
9. punctuation
10. spelling.

The criteria in the rubric, other than criterion four, are underpinned by the same conceptual elements of writing. For example, in the narrative genre, the writer should be able to orient, engage and affect the reader (ACARA 2010), and in the persuasive genre, it is the writer's capacity to orient, engage and persuade the reader (ACARA 2013). The definitions and descriptions of criteria have been adjusted to meet the requirements of the two genres and thus the two marking guides have different score ranges that reflect the progression of different written language knowledge and skills assessed by each criterion in each genre

(ACARA 2010, 2013). In either genre, the final NAPLAN writing mark is thus a summed score of the 10 criterion scores.

Importantly, NAPLAN rubrics contain an extensive set of exemplars that provide elaboration of the achievement level for each of the criteria scores and thus provide empirical reference for the NAPLAN marking rubrics. The marking rubrics and exemplars amalgamate the conceptual and evidence base on which NAPLAN writing tests are developed and used as criterion referenced assessments. New exemplars are extracted and added for each of the new NAPLAN prompts and this process is key to the development of the scoring model that is implemented in marking centres.

Marking rubrics and exemplars are also used by systems and schools to provide the teachers with diagnostic feedback on NAPLAN writing. Currently, the only information about student performance in NAPLAN writing tests is the position on the assessment scale and marking criteria scores. Jurisdictional reporting systems also provide statistical summaries and information about the distribution of marking criteria scores at a class, school, state and national level. Teachers are instructed and encouraged to use these statistical summaries to understand patterns in writing skills for their student and to inform their teaching and learning efforts and plans. Figure 1 contains an excerpt from the Victorian reporting system document (VCAA, 2013, p. 8), which represents an example of such a diagnostic use of NAPLAN marking rubrics and exemplars.



acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

Scenario 3 – Year 3 Writing

Analysis of the **Writing Criteria Report** for Year 3 identified several criteria in which students at this school were achieving generally lower scores than Year 3 students across the state or the nation.

The score distribution shown for Character and Setting revealed that the most common score for Year 3 students on this criterion is '2', both nationally and across Victoria. For this school, however, around 60% of students have a score of only '1', and the percentage of '2' scores is relatively low.

Clearly, it would be desirable to move students up from a '1' to a '2' or beyond on this criterion. Work around this criterion is likely to improve the group's ability to develop character and setting in their narrative writing and help them to move towards higher levels of achievement.

Using the NAPLAN Narrative Marking Guide

- Follow the link to the Writing Marking Guide. Look at the description for scores '1' and '2' for *Character and Setting*.
- Identify parts of the descriptors or additional information which clearly differentiate between the two scores. Note, for example, the key words that have been underlined in the marking guide extract below.

Figure 1. ***An example of diagnostic use of NAPLAN marking rubrics***

Careful and strict implementation of marking rubrics is supported by the exemplars, ensuring the validity of NAPLAN writing assessments. To enable such a valid scoring, novice markers are trained to understand and apply the marking rubrics when scoring NAPLAN scripts with and across different prompts. In addition, and equally important, to ensure the consistency of the rubrics, all markers, irrespective of their experience, must be monitored for the duration of the marking operation. These activities require concentrated efforts and are coordinated by ACARA at a national level.

It is therefore crucially important to acknowledge that the human scoring models, which are developed for each NAPLAN writing prompt, and their consistent application ensure and maintain the validity of NAPLAN writing assessments. Consequently, the statistical reliability of human scoring outcomes is fundamentally related to and is the key evidence for the validity of NAPLAN writing marking.

The pilot study on the feasibility of automated scoring in NAPLAN (ACARA, 2015a) provides preliminary empirical evidence regarding the reliability of automated scoring of NAPLAN persuasive writing. The study assessed four different scoring systems, each implementing different methods for conceptual treatment of writing, and the automated scoring result achieved. The study provided initial positive evidence on the feasibility and validity of automated scoring in NAPLAN writing, indicating that automated scoring warranted further investigation.

ACARA consequently developed a research program to collect further and more comprehensive empirical evidence on the feasibility, reliability and validity of the automated scoring in NAPLAN:

Study 1 of this research program examined a range of NAPLAN writing prompts in persuasive and narrative genres as well as scripts produced by a diverse sample of Australian students to investigate the robustness of the statistical reliability of the automated scoring in NAPLAN.

Study 2 investigated the resilience of markers and CRASE® scoring against attempts to exploit potential weaknesses of automated scoring. It also included analyses that examined other sources of empirical evidence on the validity of automated scoring.

Finally, study 3 consisted of a set of analyses that investigated some practical aspects of development of human and automated scoring models in NAPLAN, which could impact on their operational deployment.

Study 1

This study investigated empirical evidence of the feasibility and validity of automated scoring in NAPLAN, using narrative and persuasive genres, and the current NAPLAN writing test model, with separate prompts for Years 3 and 5, and Years 7 and 9. Given the importance of reliability as a necessary condition for validity, the focus of study 1 was to ascertain that automated scoring attains levels of consistency in marking equivalent to human markers, under conditions that mirror actual NAPLAN writing administration procedures.

Method

Participants and material

The NAPLAN Online scripts were collected in 2015 as part of wider NAPLAN Online transition research activities. The sample was stratified across states and territories, school types and school remoteness, and was thus broadly representative at a national level. Within each stratum, schools were nominated by the school system or were randomly chosen from the pool of schools that could administer NAPLAN online writing tests. While very remote schools were excluded from the sample for logistical reasons, ACARA worked with school authorities to include in the sample some remote schools and schools catering to students with socio-educational disadvantage. Further care was taken to ensure that the students in the sample were spread across three types of test devices: i) PCs or laptops, ii) tablets, and iii) tablets with an external keyboard. This warranted that examples of students' writing included in the study reflected the test conditions expected in the main NAPLAN Online testing event.

Table 1 provides NAPLAN 2105 school level writing mean achievement for sampled school.

Table 1. ***NAPLAN 2015 writing performance for sampled schools***

Year level	National mean	Mean	Minimum	Maximum	Number of schools
Y3	416	420	308	492	32
Y5	478	484	433	547	32
Y7	511	507	427	555	26
Y9	546	552	495	624	18

As can be seen in table 1, the study sample was broadly representative in terms of the school-level mean achievement in NAPLAN 2015 writing tests. Across year levels, the average of school means was somewhat higher than the 2015 national mean; however, in each year level the range of school performance was sufficiently wide (over one unit of standard deviation each direction).

Each student in the study completed one narrative and one persuasive prompt, with the order of prompts counterbalanced across schools. Prompts for Years 3 and 5 were administered to some Year 7 students, and some Year 5 students took the Years 7 and 9 prompts in order to provide the full range of writing scores for each prompt.

Table 2. ***Number of scripts across prompts, year levels and genre***

		Year level				
Prompt	Genre	3	5	7	9	**Total**
P1_357	Persuasive	943	656	284		**1,883**
P2_579	Persuasive		228	607	738	**1,573**
N1_357	Narrative	812	587	274		**1,673**
N2_579	Narrative		232	490	615	**1,337**
Total		**1,755**	**1,703**	**1,655**	**1,353**	**6,466**

Procedure

All the scripts in this study were scored independently by two groups of experienced NAPLAN markers. In addition, a third human score was independently awarded by a team of highly proficient and experienced NAPLAN markers, who typically serve as marking operation leaders and quality control supervisors. This third set of human scores was regarded as a 'true score' and used as a human scoring model in learning and development of automated scoring models.

Automated scoring models were independently constructed and evaluated for each of the four prompts included in this study. Within each prompt, scripts were randomly allocated to the training, validation and blind evaluation samples. The spiralling selection approach was used to ensure that approximately equal number of scripts at each score point was included in the three conditions. Consequently, each of the samples contained about one-third of the available scripts.

Scripts from the training and validation sample were used in an automated scoring model construction. The scripts from the blind evaluation sample were only used to evaluate the performance of the final scoring models.

Results

The performance of automated scoring models was evaluated using three consistency and reliability statistics:

(i) the standardised mean difference (SMD) – a statistic that provides information about the magnitude of the difference between two sets of criteria scores treated as continuous measurement

(ii) the exact agreement (EA) – an inter-marker reliability statistic that indicates how much homogeneity, or consensus, is found between the scores awarded by two sets of markers, and

(iii) the quadratic weighted kappa (QWK) – an inter-rater agreement that indicates the level of agreement between two sets of ordered scores awarded by separate markers. In addition, to measure the agreement of summed total scores, the Pearson's correlation coefficient was used.

The reliability statistics between CRASE® and a human marker, and between two sets of markers were calculated for each of the NAPLAN marking criteria. Statistics for the exact agreement and quadratic weighted kappa, as ratio, for all prompts and criteria are shown in figure 2.

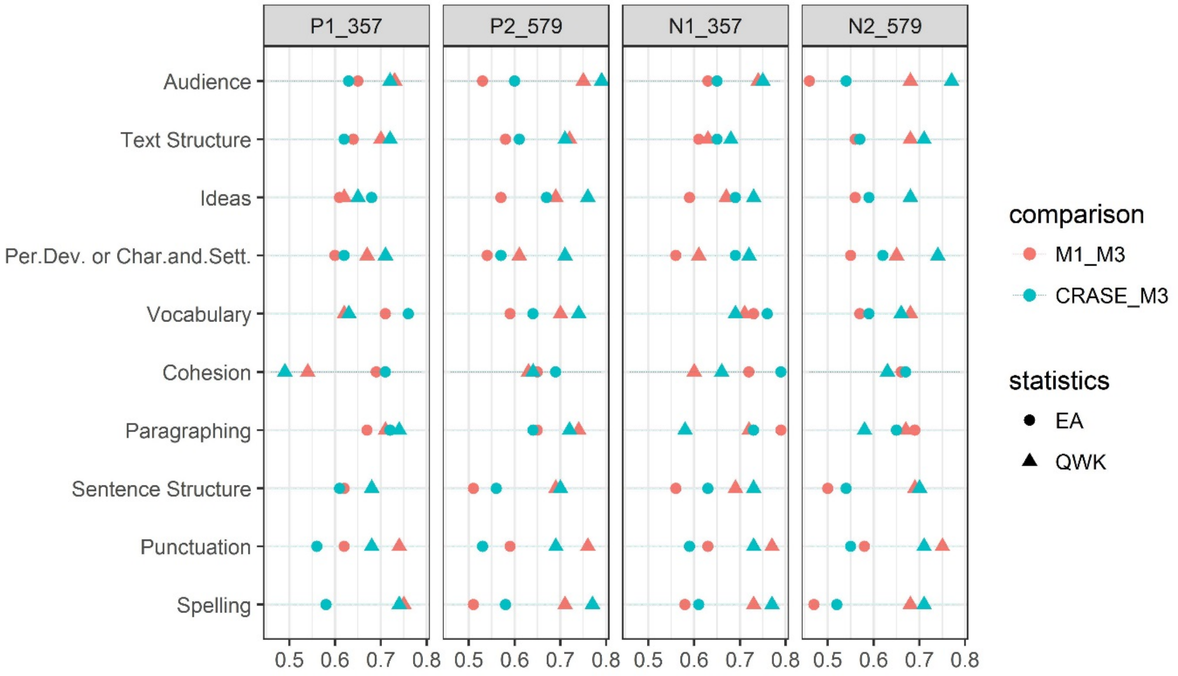


Figure 2. ***The comparison of human and automated scoring model exact agreement, as ratio, and quadratic weighted kapa statistics***

The performance of human and automated scoring models was evaluated using thresholds recommended in literature and industry practice:

(i) The absolute standardised mean difference (SMD) between CRASE® and a human marker is 0.15 or lower (Williamson, Xi, & Breyer, 2012).

(ii) The exact agreement rate (EA) between CRASE® and a human marker is 5 per cent or lower than that between two human markers (McGraw-Hill Education CTB, 2014; Pearson and ETS, 2014).

(iii) The quadratic weighted kappa (QWK) statistic case between CRASE® and a human marker is 0.10 or lower than that between two human markers (Williamson, Xi, & Breyer, 2012).

Consequently, for each of the agreements or correlation statistics, a prompt could meet evaluation threshold up to 10 times, one for each NAPLAN marking criteria. Detailed correlation and descriptive statistics for the blind validation sample are provided in appendix 1(a). A summary of the outcomes of the criteria level correlation and agreement evaluations are presented in table 3.

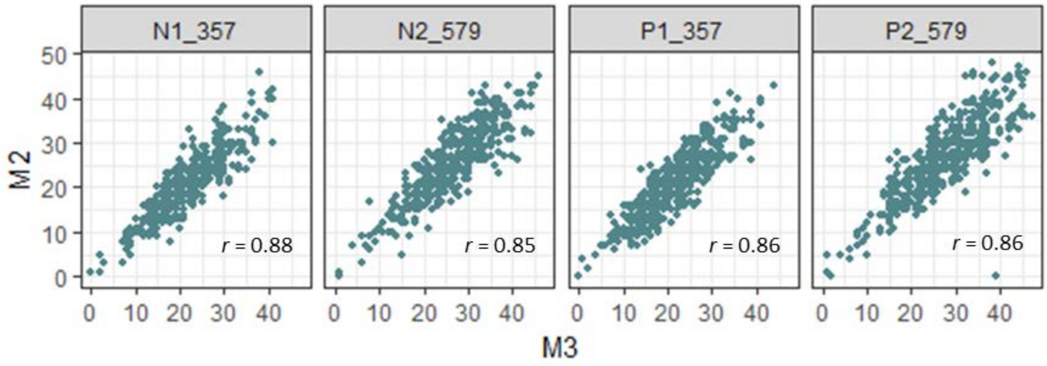
Table 3. ***Count of marking criteria scores that met the evaluation thresholds***

Task	Genre	SMD	EA	QWK
P1_357	Persuasive	10	9	10
P2_579	Persuasive	9	9	10
N1_357	Narrative	10	9	9
N2_579	Narrative	10	10	10

As shown in table 3, no prompt failed to meet all 10 evaluation thresholds in at least one statistic. Prompt N2_579 met thresholds for all marking criteria. Prompt P1_357 did not meet the exact agreement threshold for the punctuation criterion (M3–M1 = 62% vs C–M3 = 56%). The automated scoring model did not award a top score in several criteria for this prompt; however, given the relatively low number of top scores awarded by human markers, this outcome was most likely a reflection of the lack of sufficient differentiation of top scores in the human scoring model.

Prompt N1_357 did not meet standards for exact agreement and quadratic kappa statistics for the paragraphing criterion (M3–M1 = 79% vs C–M3 = 73% and M3–M1 = 0.72 vs C–M3 = 0.58 respectively). The prompt P2_579 did not meet the absolute standardised mean difference threshold for the cohesion criterion (d_{M3-C} = 0.15) and the exact agreement standard for the punctuation criterion (M3–M1 = 59% vs C–M3 = 53%). There was no pattern across prompts with regard to the criteria that were harder to mark in the automated scoring models.

The threshold for the total (summed) NAPLAN score was set as a difference of the Pearson correlation coefficients between CRASE® and a human marker of 0.05 or less than that between two human markers.



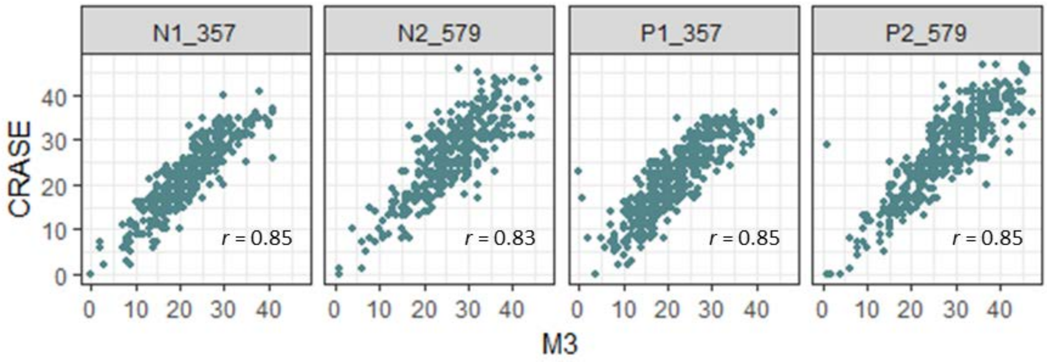


Figure 3. ***Scatterplot and Person's correlation coefficient for summed score scoring model outcome comparisons***

Figure 3 shows that in terms of the summed score, the correlation coefficient between CRASE® and human scoring was effectively the same for all prompts, and that the relationship between automated and human scoring models outcomes emulates that between two groups of human markers.

The Cronbach's α coefficient for NAPLAN 2016 writing tests is 0.95 (ACARA, 2017b). Cronbach's α coefficient is a measure of the explained variance and can be compared with the proportion of the explained variance for the correlation of two sets of independent scores. The mode correlation coefficient between two sets of human markers is *r* = 0.86, and that between human and CRASE® is *r* = 0.85. The corresponding coefficients of the explained variance are R2 = 0.93 and R2 = 0.92 respectively. These results indicate that correlation between two human scoring models and that between human and automated scoring models have come very close to the reliability level of the NAPLAN writing tasks.

The intercorrelation of criteria scores for the automated scoring models was higher than that for the human scoring models. Such an intercorrelation of criteria scores is to some extent understandable, as the same latent features space is underpinning allocation of scores for the 10 NAPLAN marking criteria. The dependency of the mark in automated scoring models, however, does not have a material impact on the distribution of summed scores. Figure 3 shows the comparison of the summed distribution of scores produced by three groups of markers and CRASE®.

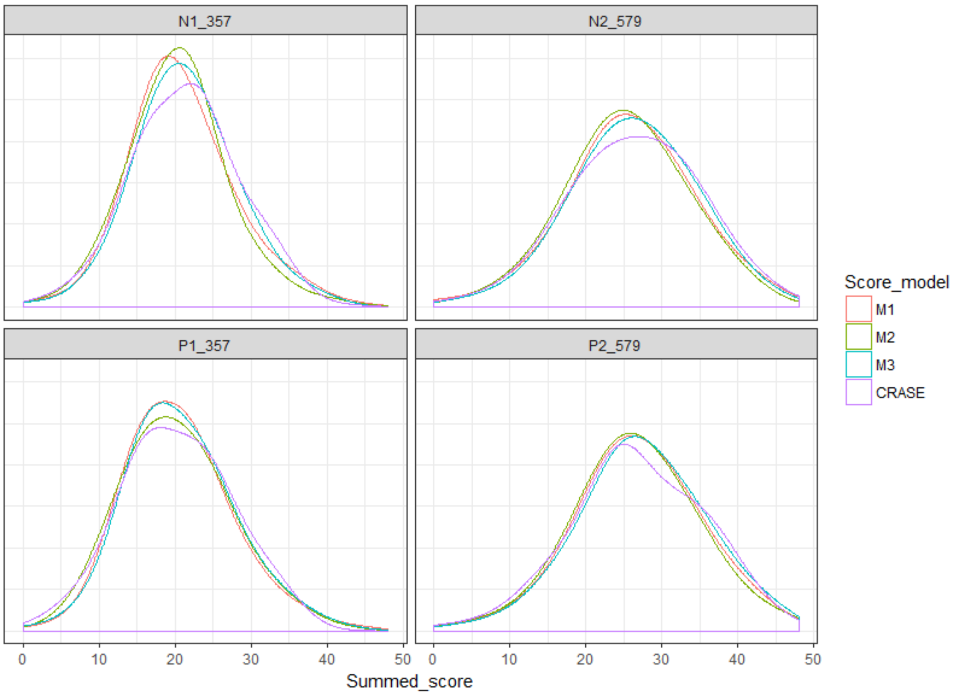


Figure 4. ***Comparison of automated and three human summed scores distributions***

As it can be seen in figure 4, the automated scoring model distribution to great extent follows that of the training human scoring model M3. Where observed, the deviation of CRASE® distribution stays with the range of deviation observed for the other two human scorings.

Discussion

The results of study 1 show that for the four NAPLAN prompts, covering both narrative and persuasive genres, automated scoring models emulated human scoring models for each of the NAPLAN marking criteria with a high level of consistency and reliability. Human markers used in this study all have high expertise and significant experience; therefore, it is reasonable to assume that their level of inter-marker reliability is higher than that found in the typical NAPLAN marking operation, where up to 30 per cent of markers are novice ones who require considerable training and monitoring.

These analyses also showed that the intercorrelation of criteria and the shape of the summed score distribution in automated scoring models requires ongoing monitoring and evaluation.

Replication analyses

Reproducibility of empirical research is recognised as a key measure of the research transparency and robustness of its findings. Consequently, evaluation of reliability and validity of automated scoring in NAPLAN writing was replicated using an additional set of eight prompts produced by a sample of students from the 2016 NAPLAN cohort, who participated in the writing test field trial.

The trial sample was a stratified random sample limited to metropolitan and provincial schools for logistic reasons. The sample is stratified across state and territories. The NAPLAN writing mean school performance was used in the sample selection to ensure that the sample had coverage across the range of writing ability, typically found in each year level. This was to ensure that the whole range of NAPLAN scores could be obtained for each trial prompt. To reduce the burden of NAPLAN Online trialling in schools, the target was limited to 200 students for each year level and each prompt.

A total of six persuasive and two narrative prompts was trialled. Each student completed two prompts, and the order of prompts was counterbalanced across schools. Table 4 shows the number of scripts for each prompt and year level.

Table 4. ***Number of scripts for trial prompts***

		Year level				
Prompt	Genre	3	5	7	9	**Total**
P3_357	Persuasive	235	201	189		**625**
P4_357	Persuasive	274	204	162		**640**
P5_357	Persuasive	279	202	137		**618**
P6_579	Persuasive		244	235	232	**711**
P7_579	Persuasive		211	233	248	**692**
P8_579	Persuasive		222	202	226	**650**
N3_357	Narrative	240	218	208		**666**
N4_579	Narrative		236	160	210	**606**
Total		**1,028**	**1,738**	**1,526**	**916**	**5,208**

Human markers found that the narrative task N4_579 did not function as expected and they found it difficult to mark. This task was, however, included in the study to assess the robustness of CRASE® marking and to ensure full transparency of the automated scoring research.

Each script was scored by two sets of highly experienced NAPLAN markers. All scores were reviewed, and scores from two markers were adjudicated if necessary. A third set of marks, consisting of adjudicated marks and existing marks, which was deemed to be the most correct, was corrected and used in the system training. Importantly, marker 1 scores were used to calculate a correlation between human and automated scoring models in the replication analyses.

In all other aspects, the evaluation of the automated scoring models followed the procedure used in the main feasibility study described. The reduction in the sample size, however, resulted in relatively sparse coverage of some scores across three different samples, which is not the most optimal situation for the construction of the automated scoring models. Detailed correlation and descriptive blind validation sample statists for eight prompts are provided in appendix 1(b). Table 5 shows the outcomes of the automated scoring evaluation for the field trial study.

Table 5. ***Count of trial prompts marking criteria scores that met the evaluation thresholds***

Prompt	Genre	SMD	EA	QWK
P3_357	Persuasive	10	8	10
P4_357	Persuasive	10	10	10
P5_357	Persuasive	10	8	10
P6_579	Persuasive	10	9	10
P7_579	Persuasive	9	8	9
P8_579	Persuasive	10	10	10
N3_357	Narrative	10	8	9
N4_579	Narrative	9	6	6

As expected, N4_579 did not pass the automated scoring evaluation for a number of criteria scores. Table 5 shows that the other persuasive and narrative automated scoring models performed well, with all but one prompt passing the standardised mean difference threshold, and all but two prompts passing the quadratic kappa threshold. Prompt N3_357 did not meet the quadratic kappa standard for paragraphing (M1–M2 = 0.82 vs C–M1 = 0.64), and the same criterion, together with the punctuation criterion, did not meet the standards for exact agreement (M1–M2 = 88% vs C–M1 = 76% and M1–M2 = 72% vs C–M1 = 66% respectively).

Prompt P7_579 did not meet the quadratic kappa standard and exact agreement for the sentence structure criterion (M1–M2 = 0.74 vs C–M1 = 0.63 and M1–M2 = 61% vs C–M1=51% respectively), and the exact agreement for the paragraphing criterion (M1–M2 = 66% vs C–M1 = 60%).

Apart from the fact that paragraphing and sentence structure were difficult to model across both genres, the analyses of human and automated scores did not yield a systematic pattern that would cause the sub-par performance of the automated scoring model for these criteria. Nonetheless, further careful examination of the automated system features and scoring models regarding these criteria is warranted.

The less than optimal conditions for model construction most affected the exact agreement statistic. Had more scripts been available for each score in the range, it is expected that these prompts would have been on par with the other two statistics. As a direct consequence of this finding and to ensure that the next round of replication studies will have a sufficient samples size, ACARA has increased the NAPLAN 2018 Online writing field trial sample to 1,000 per prompt.

All but the narrative task N4_579 described above passed the summed score Pearson’s correlation coefficient threshold for the summed score.

Summary discussion

Study 1 shows that for 11 out of 12 prompts – at both the rubric criteria and total score levels – CRASE® provides scoring outcomes that have consistency and reliability equivalent to those found between marking outcomes produced by independent groups of very experienced NAPLAN markers.

It must be noted that examples of students’ writing were not collected in the context of the main NAPLAN tests, which might have had an impact on student motivation. In addition, the sample size did not allow for the analyses of automated scoring model performance at the demographic subgroup levels. Consequently, another replication study using a sample that will enable subgroup analyses and preferably using scripts produced in the main NAPLAN Online tests, would further strengthen analyses of the validity of automated scoring in NAPLAN Online.

Study 2

In assessing the validity of automated scoring for NAPLAN writing, it is important to show validity evidence beyond the reliability results described in study 1. Research has shown that some automated scoring engines might be susceptible to artificial inflation of scores by manipulation of measurement construct-irrelevant textual features and characteristics (see Powers, Burstein, Chodorow, Fowles & Kukich, 2002; Perelman, 2012, 2014). Consequently, the robustness of automated scoring models against such deliberate manipulation of scripts was investigated in study 2. Furthermore, because the unidimensional Rasch Model was used to scale and report NAPLAN writing tests outcomes, this study also compared the latent structure of human and automated scoring models to investigate evidence of the construct validity of the automated scoring. Study 2 also includes investigation of structure of the feature space used by CRASE® to predict and assign rubric scores.

Modification and manipulation of scripts analyses

ACARA studied the existing literature and identified examples of textual manipulation that could feasibly occur under the tests administration conditions of NAPLAN writing tests (see Powers et al., 2002). To test the impact of such textual manipulation, existing NAPLAN Online essays were modified to construct 84 ‘new’ essays. The modifications consisted of a set of changes to lexical field, sentence construction, punctuation and text length. The full list is provided in appendix 2. Based on the type of the change, three sets of modified scripts were created. The expectation was that:

a. The first subset of modified scripts would receive higher than merited scores than the scores awarded to the original scripts.

b. The second subset of changes would reduce the scores of the modified scripts.

c. The third subset of changes would have no impact on the score of the modified scripts.

The modified scripts, along with their original counterparts, were inserted into the marking operation and received the same treatment by the human markers and CRASE®. In the case of CRASE® scoring, some modified scripts were included in the training sample; and some, in the blind evaluation sample. This experimental design enabled an investigation of deliberate manipulation and comparative robustness of the marking construct validity within and between human and automated scorings.

To investigate the magnitude of a potential score manipulation, the score of a modified script was compared with the score of its original. As all human essay scoring is subject to standard variability in marking, the difference of scores that were within the margin of the expected variability of NAPLAN marking (greater than four for summed scores, and greater than one for criteria scores) was deemed to be functionally equivalent. Only differences that were outside of this NAPLAN margin of variability were deemed to influence the scoring of modified scripts.

Analyses of markers and CRASE® scoring of modified scripts showed that the planned modification of the scripts did not affect the scoring of modified scripts in any systematic way, as shown in table 6:

Table 6. ***Observed score changes for three sets of modified scripts, matched automated and human scoring outcomes provided in italics***

Observed automated score change	Observed marker score change			
	Degraded	no change	improved	**Total**
	Modified with intention to degrade scores (N32)			
Degraded	*5*	2	0	**7**
no change	5	*18*	2	**25**
Improved	0	0	*0*	**0**
Total	**10**	**20**	**2**	**32**
	Modified with intention to improve scores (N46)			
Degraded	*1*	0	0	**1**
no change	2	*27*	5	**34**
Improved	1	5	5	**11**
Total	**4**	**32**	**10**	**46**
	Modified with no intention to change scores (N6)			
Degraded	*0*	0	0	**0**
no change	0	*5*	1	**6**
Improved	0	0	*0*	**0**
Total	**0**	**5**	**1**	**6**

For the scripts where modifications were expected to lead CRASE® to award higher scores and where CRASE® awarded such a score, expert markers awarded the higher score to the modified script in 5 out of 11 cases. In only one case, an expert marker awarded a score that was lower than that of the original scripts. It is also worth noting that there were five scripts that received improved scores from expert markers, but not from CRASE®. In most cases, scripts modified to degrade CRASE® scoring did not lead to reduced automated scores.

Where CRASE® awarded lower scores (5 out of 7 cases), expert markers also awarded lower scores to the modified scripts, indicating that the script modification had the same impact on automated scoring and human scoring.

These results suggest that the most likely forms of purposeful manipulation of NAPLAN scripts is not likely to produce unmerited score increases. Importantly, the same pattern of results was observed for human scoring and automated scoring, indicating that observed disturbances to the marking construct validity do not unduly impact CRASE® and automated scoring.

This research has nonetheless identified areas of script manipulation that warrant further investigation to better understand the behaviour of the CRASE® scoring algorithm. For example, when the same paragraphs were replicated in high-quality short scripts, CRASE® awarded a higher score, but the human marker did not. On the other hand, adding extra persuasive writing to a narrative prompt resulted in increases in scores by both CRASE® and human markers. These findings are being used to enhance the utilisation of a set of flags that can be triggered by such textual features. Potentially suspect essays with flags will receive further attention by expert human markers in the NAPLAN operational marking.

Dimensionality analyses

In these analyses, human and CRASE® criteria scores were subjected to principal component studies to compare scores' latent structure and to establish whether both models were underlined by a single measurement construct.

Given that the same marking rubrics were applied to all prompts and the same set of measurement model parameters were used to scale and report NAPLAN within a genre, it seemed warranted to conduct the principal component analyses using the combined responses from different prompts within the same genre. To that end, for each of the NAPLAN genre, two sets of prompt compilations were created using a different combination of prompts from the feasibility and trial studies.

The principal component analysis of human marks provided the base condition against which the dimensionality and validity of automated scores were evaluated. The key null hypothesis is that there should be no difference in the number and composition of latent structures between human and automated scoring.

Separate principal component analyses for the four compilation of scripts showed no difference in the latent structure between human and AES scores. In all cases, a singe principal component had only one eigenvalue higher than one; thus, the same unidimensional latent structere suffiicently explained the variance of the 10 critera scores for both human and automated scoring model outcomes, as shown in table 7.

Table 7. ***The proportion of the explain variance for the first principal component in human and automated scoring models***

	Narrative		Persuasive	
	Compilation 357	Compilation 579	Compilation 357	Compilation 579
Marker scores	67.6%	71.3%	72.1%	74.6%
CRASE scores	79.7%	82.6%	77.0%	85.3%

Furthermore, the analyses show that in both the human and automated scoring data, the proportions of explained variance for the other nine principal components have the same pattern and magnitude, as shown in figure 4.

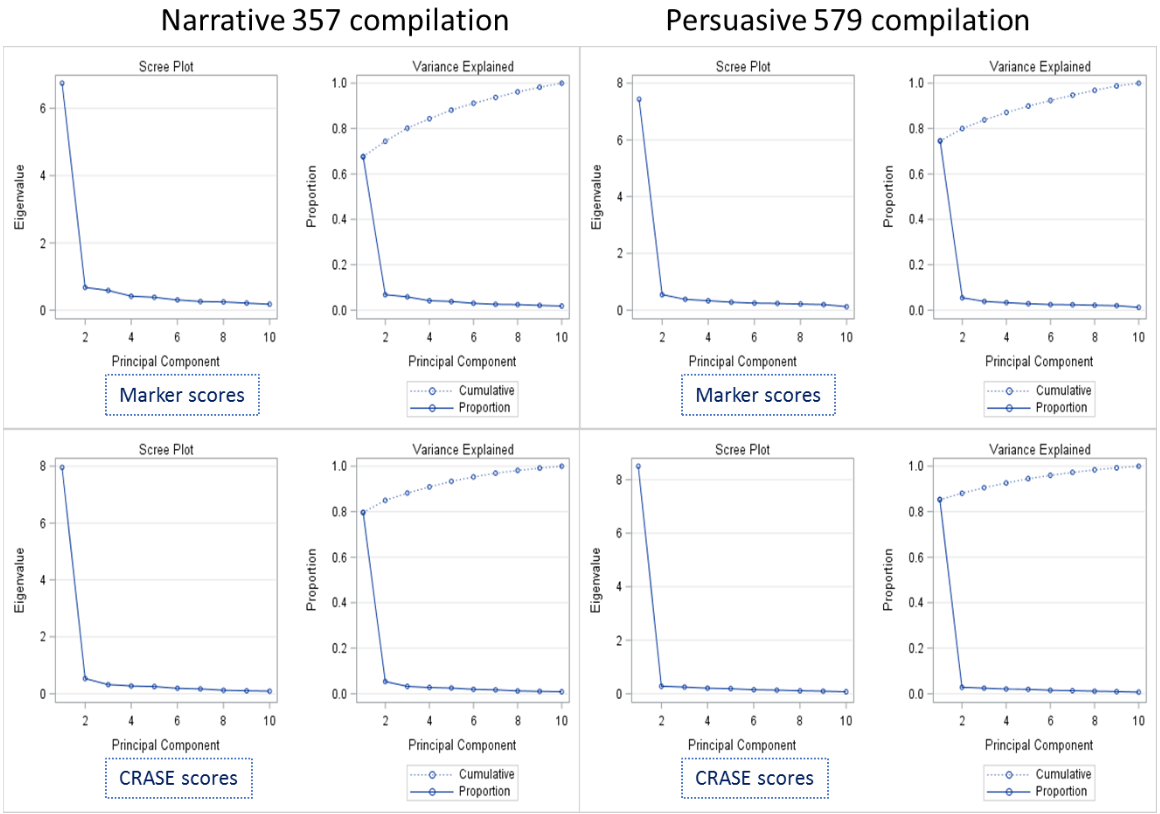


Figure 4. ***Principal component analyses' results for human and automated scoring models***

These analyses show that there is no material difference in latent structure between human and CRASE® criteria scores and that a single principal component is sufficient to explain the score variance in both cases.

Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) refer to unidimensionality of a test as validity evidence based on the internal structure of a measurement instrument. The results of the dimensionality analyses show that automated scoring models have the same internal structure as the human scoring models.

Latent structure of automated scoring system’s features

CRASE® uses natural language processing methods to extract a set of quantifiable features as a key step in building of each automated scoring model. These features are then submitted to the machine learning modelling and used to predict and award a score separately for each NAPLAN marking criterion. The description of features and their exact nature is intellectual property and commercially confidential information. However, these automated system features can be treated as observed, and the internal relationship between such features or their latent structure can still be investigated. ACARA designed and commissioned such an investigation as it can provide important information about the response process implemented in the automated scoring of NAPLAN writing.

A data set containing the loadings on the 26 features CRASE® used to construct NAPLAN or any other scoring model was constructed for each of the four prompt compilations used in the principal component analyses described in this paper.

Cattell's scree test (Cattell, 1966) inspects the shape of the eigenvalues curve in order to detect the point at which the curve changes significantly. The scree tests show that across the four data sets, up to 10 principal components are needed to explain the latent variance of the CRASE® ‘s features (see figure 5). The orthogonal principal component analyses suggest no correlation between the components; thus, to investigate whether further reduction of the latent space structures was possible, oblique principal component analysis, which allowed correlations between components, was conducted.

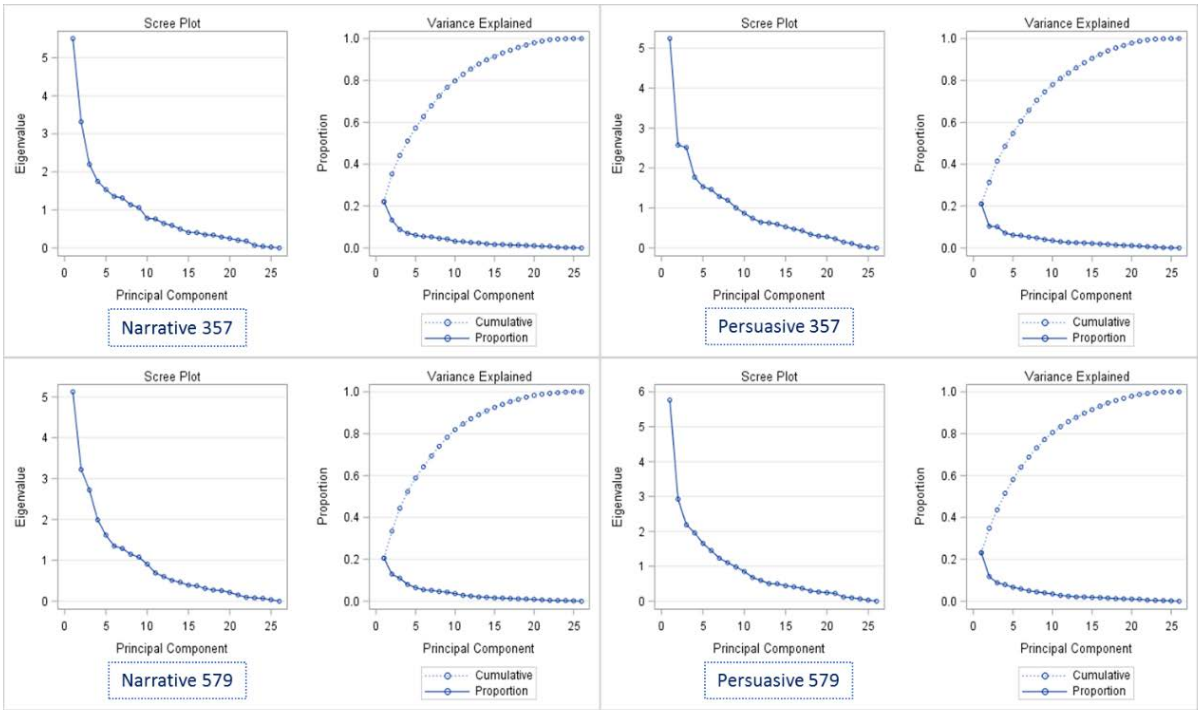


Figure 5. ***Principal component analyses' results for CRASE® features data***

The oblique principal component analyses show: there are five latent dimensions that sufficiently explain the latent variance of the feature space CRASE® uses to predict the final score. These five dimensions can be projected onto CRASE®'s features that capture lexical, syntactic and semantic aspects of writing. Some difference in the composition of the five extracted latent dimensions between different genres was also observed, providing evidence that CRASE®'s features can generate a different representation of narrative and persuasive writing.

The dimensionality analyses of the automated scoring system's features show that internal structure of CRASE features can detect different elements of the written language. This means that CRASE® can:

- extract from the scripts and human scoring data patterns and information about elements of the written language, which are most relevant for different NAPLAN marking criteria and
- construct adequate scoring models for each criterion and its scoring range.

Study 3

Study 3 analyses were developed and implemented to address the practical aspects of operational deployment of automated scoring in NAPLAN in relation to:

- characteristics of human scoring used in the development of automated scoring models
- capability of automated scoring system to flag aberrant and disturbing content scripts
- relationship between criteria score thresholds produced by the automated scoring and those of NAPLAN writing scales.

Variability of the human scoring

In automated scoring research, standard practice is to provide two sets of human marks. The choice of a mark has been identified as an important factor in the construction of automated scoring models and their evaluation (see Perelman, 2014, and Shermis, 2014). Consistent with this approach, the set of four essays used in study 1 was marked by two teams of experienced NAPLAN markers and by a small group of expert markers. (These expert markers typically served as NAPLAN marking centre supervisors and leaders.) The expert markers scored all essays without access to the other two sets of marks; their marks were treated as the 'true' mark for each essay. The data showed that the correlation between the 'true' mark and any of the other two human scores was not significantly different from that between the latter two group of markers; for example, three correlation coefficients for the summed score on a persuasive task were r = 0.86, r = 0.87 and r = 0.84 respectively. Moreover, the distribution of scores across three sets of human scores, as shown in figure 2, is very similar and thus, the automated scoring models were not affected in any material way by the selection of the human score used their development.

In the replication analyses, two independent sets of marks were adjudicated. Any score that was deemed to be inadequate was adjusted to produce an accurate mark. The final set of marks, used in the system training, consisted of these adjudicated marks and existing marks that were deemed to be the most correct. Such a treatment of writing scores surpasses the level of control and consistency of current NAPLAN marking operations.

Further differences between marking operations for the two studies discussed above were that one was conducted using the distributed marking model, where markers worked from home; and the other, where markers were located in a marking centre. Such a difference in the marking operation may potentially impact on the level of consistency and reliability of marking; consequently, inter-marker reliability within and across two marking centre setups was investigated.

In these analyses, the generalizability theory (G theory) was used to evaluate the magnitude of the marker variability across two marking centre types.

The narrative data from the study 1 produced a marker main effect variance component of 0.71. The narrative data from the replication analyses showed a marker nested effect variance component of 1.41. When these variance components were divided by the sum of the variance components within their respective studies, percentages of 1.1 and 2.3 were obtained. The persuasive data from the study 1 produced a marker main effect variance component of 2.03, and the replication analyses that of 0.63. When these variance components were divided by the sum of the variance components within their respective studies, percentages of 2.7 and 1.1 were obtained. The percentages were somewhat reversed between compared to centre conditions, but the observed marker variability was equally low. The evidence suggests that there is very little difference in the variability of markers across genres and marking centre conditions.

Aberrant and disturbing content flags

The capability of an automated scoring system to recognise and flag scripts, which require human attention and evaluation, is important for both the validity of automated scoring and for its operational deployment in the marking of student scripts. Scripts used in the feasibility studies have been scored by human markers and have received an aberrant script flag according to the NAPLAN rubric guide. Consequently, the ability of the automated scoring system to recognise such scripts was investigated.

The markers were able to flag scripts based on one or more characters of the script, and a script could receive more than one flag. A review of these flags and their corresponding colour codes indicate the following:

(i) Student has included a plan, presumed to mean that the student included an outline or planning materials as part of their script.
(ii) Student's script is off task, often meaning that a narrative script was submitted for a persuasive writing task or vice versa.
(iii) A catch-all category that includes plagiarism, blocks of repeated text, non-serious attempts, and so on.
(iv) Scripts that contain unusual characters.

The number of scripts that CRASE® flagged for each of these categories were compared with the flags assigned by human markers. The purpose of these analyses was to collect data to inform the setting of statistical thresholds that the automated scoring system could use to determine aberrant scripts in the operational marking.

In collaboration with the Pacific Metric research team, ACARA submitted the scripts from the feasibly study to the automated content analyses for potential abuse and self-harm that students could include in their writing. This and other disturbing content is flagged by human markers as a standard practice in the NAPLAN marking. A prototype of a disturbing content filter, developed by Pacific Metrics, was deployed and the outcomes of its deployment were compared with the disturbing content flags raised by human markers. The NAPLAN writing tests manager conducted detailed analyses of scripts that were flagged by human markers and the system. The initial results and the prototype of the disturbing content filter show a capacity to flag most of the scripts that were flagged by human markers; however, this is an area of development that needs further research and more empirical data.

Correspondence of the NAPLAN marking rubrics thresholds

NAPLAN writing criteria scores are discrete points on an ordinal scale. To replicate such scores, an automated scoring system produces a continuous scale that is used to determine the final discrete scores (ACARA, 2015b). This process reverses the psychometric analyses of NAPLAN writing, where ordinal scale scores are transformed into the continuous scale thresholds using the Partial Credit Model. Common to these two processes is that they produce a continuous scale on which criteria scores are represented as thresholds and that both scales are products of the latent variance analyses. Thus, it is possible to compare CRASE®'s criteria thresholds, which are projections of the latent scoring features structure, with those used to scale and report students' performance in NAPLAN.

The NAPLAN writing latent thresholds are highly stable across different prompts and test administrations – so much that the same set of thresholds has been used in the scaling of NAPLAN writing since 2011 (ACARA, 2015b).

Cross-prompt genre scoring models were used to extract a single set of CRASE®'s continuous scale thresholds for each of the four automated scoring models. These thresholds were compared with sets of the current NAPLAN writing scale thresholds. Each of the correlations between the NAPLAN writing scale thresholds and automated score models' parameters range of 0.95–0.99, indicating highly linear, positive relationships. It should be noted that the relationship between the automated scoring model and NAPLAN scale thresholds for the spelling criterion has the same high level of correlation and linear relationship, but follows a somewhat different pattern to that of other criteria. The most likely reason for this different pattern is that human markers had permission and the ability to differentiate spelling errors from typographical errors. For example, ‘cst’ is considered to be a typographical error by human markers, as the key for the letter ‘s’ is close to the key for the letter ‘a’, and thus it could be assumed that a student intended to type the word ‘cat’. In contrast, the automated scoring system simply counts it as a spelling error. It is worth noting though that this observation seems to be limited to the cross-prompt scoring models. The task-based automated scoring models consistently and reliably score the spelling marking criterion.

The results indicate concordance between measurement parameters used in constructing automated scoring models and the assessment scale parameters for persuasive and narrative NAPLAN writing, and thus provide additional empirical evidence regarding the validity of the internal structures of automated scoring model outcomes at the genre level (AERA, APA, & NCME, 2014).

Summary discussion

The NAPLAN Online automated essay scoring research program was designed to investigate and collect evidence on the feasibility and validity of automated scoring in NAPLAN writing. The research focuses on the key aspects of the NAPLAN writing test itself; on the understanding of the key writing skills and knowledge, expressed in the Australian Curriculum; and on whether automated scoring can successfully cater for both aspects in the NAPLAN writing assessments.

As this is applied research, the interpretation of results is restricted to the context of NAPLAN Online writing assessment.

Study 1 and its replication analyses provided empirical evidence on the validity of automated scoring in NAPLAN writing assessments in relation to the validity of tests content and response processes (AERA, APA, & NCME, 2014).

Further empirical evidence on the validity of automated scoring was provided in study 2. CRASE®'s showed promising evidence in its ability to recognise attempts to manipulate its marking. Results also show that the latent structure of automated scores is the same as that of the human markers. Finally, there is preliminary evidence that the internal CRASE® features are able to identify and process key elements of written language.

Study 3 analyses show no operational impediments to implantation of automated scoring in NAPLAN.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC, American Psychological Association.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2010). *Writing Narrative Marking Guide.* Retrieved from: www.nap.edu.au/_resources/2010_Marking_Guide.pdf

Australian Curriculum, Assessment and Reporting Authority(ACARA). (2013). *Persuasive Writing Marking Guide*. Retrieved from www.nap.edu.au/_resources/Amended_2013_Persuasive_Writing_Marking_Guide_-With_cover.pdf

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2015a). *Evaluation of Automated Scoring of NAPLAN Persuasive Writing*. Retrieved from www.nap.edu.au/online-assessment/research-and-development/automated-essay-scoring

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2015b). *NAPLAN 2014 Technical Report.* Retrieved from www.nap.edu.au/_resources/2014_NAPLAN_technical_report.pdf

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2017a). *The Australian National Assessment Program Literacy and Numeracy (NAPLAN) assessment framework: NAPLAN Online 2017-2018*. Retrieved from www.nap.edu.au/docs/default-source/default-document-library/naplan-assessment-framework.pdf?sfvrsn=2

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2017b). *NAPLAN 2016 Technical Report.* Retrieved from www.nap.edu.au/docs/default-source/default-document-library/2016-naplan-technical-report.pdf?sfvrsn=2

Cattell R. B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–76.

Haswell, R., & Wilson, M. (2013). Professionals against machine scoring of student essays in high-stakes assessment. Retrieved July 11, 2016 from http://humanreaders.org/petition/index.php

Lochbaum, K.E., Foltz, P.W., Mao, X., Tong, Y, Way, W., Hanlin, S.M., Rosenstein, M., Burstein, J., Cahill, A., Heilman, M., Lorenz, F., Zhang, M., & Rupp, A.A. (2015). Research results of PARCC automated scoring proof of concept study. Retrieved from http://parcc-assessment.org/images/Resources/Educator-resources/PARCC_AI_Research_Report.pdf

McGraw-Hill Education CTB (2014). *Smarter Balanced Assessment Consortium Field Test: Automated Scoring Research Studies (in accordance with Smarter Balanced RFP 17).* Retrieved from www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf

Pearson and Educational Testing Service (2015). *Research Results of PARCC Automated Scoring Proof of Concept Study.* Retrieved from www.parcconline.org/images/Resources/Educator-resources/PARCC_AI_Research_Report.pdf

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In A. Bazerman, C; Dean, C; Early, J; Lunsford, K; Null, S; Rogers, P; Stansell (Ed*.), International advances in writing research* (pp. 121–31). Fort Collins, CO: The WAC Clearinghouse and Parlor Press.

Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hamner, "Contrasting state-of-the-art automated scoring of essays: Analysis". The Journal of Writing Assessment, 6, 1.

Perelman, L. (2014). When "the state of the art" is counting words. Assessing. *Writing, 21*, 104–11.

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2002). Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior, 18(2)* 103–34.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.

Victorian Curriculum and Assessment Authority (VCAA). (2013). *Using NAPLAN data diagnostically: the introductory guide for classroom teachers.* Retrieved from www.vcaa.vic.edu.au/Documents/naplan/teachersguide-usingnaplandata.pdf

Williamson, D., Xi, X., & Breyer, F. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(1),* 2–13.

Appendix 1

Appendix 1a. Study 1: Blind validation sample results

Criteria scores correlations

The criteria-level analyses appearing in this section of the report outline the results on the blind validation sample and examine CRASE relative to the three markers. For each prompt and criteria, the following statistics are presented: (1) the number of scores for each marker and CRASE, (2) the percent of students at score point for each marker and CRASE, (3) the mean scores and standard deviations for each marker and CRASE, (4) the standardised mean difference between marker 3 and CRASE using the pooled standard deviation, (5) the exact, adjacent, and non-adjacent rates for marker 1 and marker 2, marker 1 and CRASE, marker 1 and marker 3, and CRASE and marker 3, and (6) kappa, quadratic weighted kappa (QWK), and the Pearson correlation for the four comparisons.

Because of the large number of criteria (10), each prompt has four tables presenting the statistics. Shading is used to identify these conditions:

- The scoring source assigned no scores at a score point.
- The absolute standardised mean difference was 0.15 or greater.
- The CRASE – marker 1 exact agreement rate was 5 per cent or lower than the marker 1 – marker 2 exact agreement rate.
- The CRASE – marker 3 exact agreement rate was 5% or lower than the marker 1 – marker 3 exact agreement rate.
- The quadratic weighted kappa statistic for CRASE and marker 1 was 0.10 or lower than the marker 1 – marker 2 QWK.
- The quadratic weighted kappa statistic for CRASE and marker 3 was 0.10 or lower than the marker 1 – marker 3 QWK.
- The correlation between CRASE and marker 1 was 0.10 or lower than the marker 1 – marker 2 correlation.
- The correlation between CRASE and marker 3 was 0.10 or lower than the marker 1 – marker 3 correlation.

Prompt P1_357

Table 1. Marker and CRASE agreement for audience, text structure and ideas for P1_357

	Audience				Text Structure				Ideas			
	M1	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**
N	614	614	615	615	614	614	615	615	614	614	615	615
Score Distributions												
0	0.50%	0.20%	0.30%	0.20%	6%	7%	7%	7%	1%	1%	0.80%	1%
1	5%	9%	6%	8%	38%	43%	40%	41%	11%	15%	8%	10%
2	47%	45%	47%	45%	42%	37%	37%	38%	51%	46%	46%	45%
3	36%	35%	33%	34%	11%	10%	15%	14%	31%	33%	42%	42%
4	9%	9%	11%	12%	3%	2%	0.80%	**0.00%**	5%	4%	4%	2%
5	2%	1%	2%	**0.00%**					0.70%	**0.00%**	0.20%	0.20%
6	0.20%	0.20%	**0.00%**	**0.00%**								
Mean	2.55	2.48	2.56	2.5	1.67	1.56	1.63	1.58	2.29	2.24	2.4	2.33
SD	0.85	0.85	0.86	0.82	0.85	0.84	0.84	0.82	0.8	0.8	0.73	0.74
d_{M3-C}				0.07				0.06				0.1
Agreement Indices												
	M1–M2	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**
Exact	62%	63%	65%	63%	64%	64%	64%	62%	59%	63%	61%	68%
Adjacent	36%	36%	33%	36%	35%	35%	35%	37%	39%	34%	37%	31%
Non-Adjacent	2%	1%	2%	1%	1%	2%	2%	0.50%	2%	3%	2%	1%
Kappa	0.41	0.43	0.46	0.44	0.46	0.45	0.45	0.44	0.36	0.41	0.38	0.47
QWK	0.69	0.71	0.73	0.72	0.73	0.7	0.7	0.72	0.62	0.57	0.62	0.65
Pearson *r*	0.7	0.71	0.73	0.72	0.73	0.71	0.7	0.72	0.62	0.58	0.63	0.66

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 2. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P1_357

	Persuasive Devices				Vocabulary				Cohesion			
	M1	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**
N	614	614	615	615	614	614	615	615	614	614	615	615
Score Distributions												
0	9%	8%	7%	8%	0.50%	0.30%	0.80%	0.80%	1%	0.80%	1%	1%
1	42%	40%	39%	39%	6%	8%	5%	5%	17%	24%	15%	18%
2	38%	40%	40%	41%	64%	65%	72%	70%	68%	64%	74%	72%
3	9%	11%	11%	11%	25%	22%	18%	23%	13%	10%	10%	10%
4	2%	2%	2%	**0.00%**	4%	4%	4%	0.50%	0.70%	0.50%	0.80%	**0.00%**
5					0.50%	0.20%	0.30%	**0.00%**				
6												
Mean	1.54	1.59	1.61	1.56	2.28	2.21	2.2	2.17	1.94	1.85	1.95	1.9
SD	0.87	0.84	0.84	0.8	0.68	0.67	0.64	0.55	0.61	0.62	0.56	0.55
d_{M3-C}				0.06				0.05				0.09
Agreement Indices												
	M1–M2	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**
Exact	56%	59%	60%	62%	68%	72%	71%	76%	65%	71%	69%	71%
Adjacent	41%	38%	38%	37%	31%	26%	28%	23%	34%	28%	30%	29%
Non-Adjacent	3%	3%	3%	0.70%	0.80%	1%	1%	0.70%	0.50%	0.80%	0.30%	0.50%
Kappa	0.33	0.38	0.39	0.43	0.39	0.44	0.4	0.47	0.32	0.38	0.34	0.33
QWK	0.63	0.65	0.67	0.71	0.62	0.59	0.62	0.63	0.52	0.52	0.54	0.49
Pearson *r*	0.63	0.65	0.67	0.71	0.62	0.61	0.62	0.63	0.53	0.53	0.54	0.5

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 3. Marker and CRASE agreement for paragraphing, sentence structure, and punctuation P1_357

	Paragraphing				Sentence Structure				Punctuation			
	M1	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**
N	614	614	615	615	614	614	615	615	614	614	615	615
Score Distributions												
0	47%	42%	43%	43%	0.80%	0.70%	0.80%	1%	5%	4%	3%	5%
1	31%	37%	41%	40%	9%	11%	8%	10%	23%	21%	23%	23%
2	18%	20%	15%	17%	45%	46%	45%	43%	38%	39%	40%	39%
3	3%	1%	1%	**0.00%**	36%	33%	34%	37%	27%	29%	28%	31%
4					8%	9%	11%	9%	6%	7%	5%	3%
5					1%	0.80%	1%	**0.00%**	0.80%	0.30%	**0.00%**	0.20%
6					0.20%	**0.00%**	**0.00%**	0.20%				
Mean	0.77	0.8	0.75	0.73	2.46	2.41	2.5	2.44	2.1	2.15	2.07	2.05
SD	0.85	0.79	0.75	0.73	0.86	0.85	0.87	0.84	1	0.97	0.92	0.91
$d_{M3\text{-}C}$				0.03				0.07				0.02
Agreement Indices												
	M1–M2	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**
Exact	63%	66%	67%	72%	61%	60%	62%	61%	58%	**52%**	62%	**56%**
Adjacent	35%	33%	31%	28%	37%	38%	35%	37%	39%	44%	35%	42%
Non-Adjacent	2%	1%	2%	0.20%	2%	2%	3%	2%	3%	5%	3%	3%
Kappa	0.43	0.47	0.49	0.55	0.4	0.4	0.43	0.41	0.41	0.32	0.46	0.37
QWK	0.68	0.69	0.71	0.74	0.68	0.67	0.68	0.68	0.74	0.66	0.74	0.68
Pearson *r*	0.68	0.7	0.71	0.74	0.68	0.67	0.68	0.68	0.74	0.66	0.74	0.68

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 4. Marker and CRASE agreement for spelling for P1_357

	Spelling			
	M1	**M2**	**M3**	**C**
N	614	614	615	615
Score Distributions				
0	0.50%	0.20%	0.50%	0.30%
1	5%	7%	4%	3%
2	30%	28%	29%	28%
3	40%	38%	36%	37%
4	21%	20%	23%	26%
5	4%	6%	8%	5%
6	0.30%	0.80%	0.30%	0.20%
Mean	2.89	2.93	3.03	3
SD	0.96	1.04	1.03	0.96
$d_{M3\text{-}C}$				0.03
Agreement Indices				
	M1–M2	**M1–C**	**M1–M3**	**C–M3**
Exact	54%	62%	58%	58%
Adjacent	42%	36%	40%	40%
Non-Adjacent	4%	2%	3%	2%
Kappa	0.36	0.46	0.41	0.42
QWK	0.71	0.75	0.75	0.74
Pearson *r*	0.71	0.75	0.76	0.74

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Prompt N1_357

Table 5. Marker and CRASE agreement for audience, text structure, and ideas for N1_357

	Audience				Text Structure				Ideas			
	M1	M2	M3	C	M1	M2	M3	C	M1	M2	M3	C
N	546	546	545	547	546	546	545	547	546	546	545	547
Score Distributions												
0	0.00%	1%	0.20%	0.20%	4%	3%	3%	3%	1%	1%	0.90%	1%
1	5%	8%	6%	5%	34%	35%	31%	30%	18%	21%	10%	8%
2	54%	47%	43%	43%	48%	50%	46%	48%	40%	43%	38%	38%
3	28%	33%	36%	37%	12%	10%	17%	19%	33%	31%	44%	47%
4	9%	8%	11%	13%	2%	1%	2%	0.20%	6%	4%	7%	5%
5	4%	3%	3%	1%					1%	0.70%	0.40%	0.20%
6	0.70%	0.40%	0.60%	0.20%								
Mean	2.55	2.48	2.66	2.63	1.76	1.71	1.82	1.84	2.29	2.18	2.47	2.47
SD	0.93	0.92	0.92	0.84	0.8	0.75	0.81	0.76	0.92	0.86	0.82	0.78
d_{M3-C}				0.03				-0.03				0
Agreement Indices												
	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3
Exact	60%	63%	63%	65%	63%	61%	61%	65%	54%	60%	59%	69%
Adjacent	35%	36%	34%	34%	35%	38%	37%	33%	41%	37%	38%	29%
Non-Adjacent	5%	1%	2%	0.70%	2%	0.90%	2%	1%	5%	3%	3%	1%
Kappa	0.38	0.43	0.44	0.48	0.4	0.39	0.39	0.46	0.33	0.41	0.4	0.52
QWK	0.62	0.74	0.74	0.75	0.62	0.66	0.63	0.68	0.61	0.67	0.67	0.73
Pearson *r*	0.62	0.74	0.75	0.76	0.62	0.66	0.63	0.68	0.61	0.69	0.69	0.73

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 1. Marker and CRASE agreement for character and setting, vocabulary and cohesion for N1_357

	Character and Setting				Vocabulary				Cohesion			
	M1	M2	M3	C	M1	M2	M3	C	M1	M2	M3	C
N	546	546	545	547	546	546	545	547	546	546	545	547
Score Distributions												
0	3%	5%	2%	2%	0.90%	1%	0.70%	0.70%	1%	0.90%	1%	1%
1	33%	31%	26%	25%	6%	8%	6%	6%	16%	19%	13%	11%
2	44%	48%	49%	50%	60%	66%	67%	67%	70%	68%	73%	76%
3	18%	14%	21%	23%	28%	21%	21%	24%	11%	11%	11%	11%
4	3%	2%	2%	0.50%	5%	4%	4%	2%	1%	0.40%	2%	0.50%
5					0.50%	0.50%	0.60%	0.20%				
6												
Mean	1.85	1.77	1.95	1.95	2.32	2.2	2.24	2.21	1.96	1.91	1.99	1.99
SD	0.84	0.82	0.79	0.76	0.72	0.7	0.68	0.61	0.61	0.59	0.6	0.53
d_{M3-C}				0				0.05				0
Agreement Indices												
	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3
Exact	54%	61%	56%	69%	70%	73%	73%	76%	71%	74%	72%	79%
Adjacent	44%	38%	41%	30%	28%	26%	27%	22%	28%	25%	28%	21%
Non-Adjacent	2%	0.90%	2%	0.70%	1%	0.70%	0.40%	0.70%	0.70%	0.40%	0.40%	0.20%
Kappa	0.3	0.4	0.34	0.51	0.45	0.49	0.49	0.52	0.39	0.4	0.38	0.5
QWK	0.61	0.67	0.61	0.72	0.68	0.68	0.71	0.69	0.56	0.59	0.6	0.66
Pearson *r*	0.61	0.68	0.62	0.72	0.69	0.7	0.72	0.7	0.56	0.6	0.6	0.67

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 2. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for N1_357

	Paragraphing				Sentence Structure				Punctuation			
	M1	M2	M3	C	M1	M2	M3	C	M1	M2	M3	C
N	546	546	545	547	546	546	545	547	546	546	545	547
Score Distributions												
0	49%	48%	48%	44%	1%	1%	0.90%	1%	5%	5%	4%	4%
1	42%	45%	46%	51%	7%	9%	7%	7%	23%	22%	24%	25%
2	9%	7%	6%	5%	43%	41%	37%	35%	39%	38%	41%	37%
3					35%	40%	41%	42%	26%	30%	27%	27%
4					11%	8%	12%	13%	7%	5%	4%	6%
5					2%	1%	2%	0.20%	0.50%	0.40%	**0.00%**	0.70%
6					0.40%	0.20%	**0.00%**	0.40%				
Mean	0.6	0.59	0.59	0.61	2.54	2.49	2.62	2.6	2.08	2.1	2.02	2.07
SD	0.65	0.62	0.61	0.58	0.93	0.87	0.87	0.89	1	0.96	0.91	0.99
d_{M3-C}				-0.03				0.02				-0.05
Agreement Indices												
	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3
Exact	74%	70%	79%	**73%**	59%	58%	56%	63%	61%	**56%**	63%	59%
Adjacent	25%	29%	20%	26%	38%	41%	42%	35%	38%	41%	35%	39%
Non-Adjacent	0.90%	0.70%	0.40%	0.70%	3%	2%	2%	2%	2%	4%	2%	2%
Kappa	0.55	0.47	0.64	0.51	0.38	0.38	0.35	0.45	0.45	0.39	0.48	0.42
QWK	0.65	0.57	0.72	**0.58**	0.68	0.71	0.69	0.73	0.77	0.71	0.77	0.73
Pearson *r*	0.65	0.58	0.73	**0.58**	0.68	0.71	0.69	0.73	0.77	0.71	0.77	0.73

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 3. Marker and CRASE agreement for spelling for N1_357

	Spelling			
	M1	M2	M3	C
N	546	546	545	547
Score Distributions				
0	0.20%	0.60%	0.40%	0.40%
1	3%	5%	4%	5%
2	24%	21%	22%	22%
3	44%	48%	42%	44%
4	23%	20%	26%	24%
5	6%	4%	6%	5%
6	0.40%	0.90%	0.20%	0.20%
Mean	3.05	2.97	3.09	3.01
SD	0.93	0.96	0.96	0.94
d_{M3-C}				0.08
Agreement Indices				
	M1–M2	M1–C	M1–M3	C–M3
Exact	59%	60%	58%	61%
Adjacent	39%	39%	40%	38%
Non-Adjacent	2%	0.90%	2%	0.70%
Kappa	0.4	0.43	0.4	0.45
QWK	0.73	0.76	0.73	0.77
Pearson *r*	0.73	0.76	0.74	0.78

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Prompt N2_579

Table 4. Marker and CRASE agreement for audience, text structure, and ideas for N2_579

	Audience				Text Structure				Ideas			
	M1	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**
N	435	433	434	434	435	433	434	434	435	433	434	434
Score Distributions												
0	3%	0.20%	0.20%	0.90%	3%	1%	3%	3%	2%	1%	1%	0.70%
1	4%	3%	3%	3%	15%	17%	15%	14%	8%	12%	5%	6%
2	18%	26%	21%	20%	44%	46%	39%	36%	26%	26%	19%	21%
3	37%	35%	34%	31%	28%	30%	36%	39%	42%	39%	53%	52%
4	25%	24%	27%	30%	10%	6%	8%	8%	17%	19%	17%	16%
5	11%	9%	12%	12%					5%	3%	4%	4%
6	3%	2%	3%	3%								
Mean	3.23	3.14	3.33	3.33	2.28	2.23	2.32	2.35	2.81	2.72	2.93	2.9
SD	1.22	1.09	1.13	1.15	0.94	0.85	0.91	0.91	1.04	1.05	0.93	0.91
$d_{M3\text{-}C}$				0				-0.03				0.03
Agreement Indices												
	M1–M2	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**
Exact	48%	52%	46%	54%	59%	61%	56%	57%	45%	60%	56%	59%
Adjacent	44%	42%	46%	41%	38%	38%	40%	42%	48%	36%	39%	37%
Non-Adjacent	8%	7%	8%	4%	3%	1%	3%	1%	7%	4%	5%	4%
Kappa	0.31	0.36	0.28	0.4	0.4	0.44	0.37	0.37	0.25	0.42	0.37	0.37
QWK	0.67	0.72	0.68	0.77	0.68	0.75	0.68	0.71	0.64	0.71	0.68	0.68
Pearson *r*	0.68	0.72	0.68	0.77	0.69	0.75	0.68	0.71	0.64	0.72	0.69	0.68

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 5. Marker and CRASE agreement for character and setting, vocabulary, and cohesion for N2_579

	Character and Setting				Vocabulary				C ohesion			
	M1	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**
N	435	433	434	434	435	433	434	434	435	433	434	434
Score Distributions												
0	2%	2%	2%	2%	1%	0.70%	0.70%	1%	2%	1%	0.90%	1%
1	12%	14%	10%	8%	3%	5%	3%	2%	9%	11%	7%	5%
2	45%	40%	39%	38%	37%	42%	41%	41%	57%	59%	58%	55%
3	32%	35%	38%	40%	40%	36%	37%	38%	26%	26%	30%	34%
4	10%	9%	10%	11%	15%	14%	15%	16%	6%	3%	3%	4%
5					3%	2%	3%	0.90%				
6												
Mean	2.35	2.36	2.44	2.49	2.73	2.64	2.72	2.68	2.24	2.18	2.28	2.34
SD	0.88	0.9	0.88	0.88	0.92	0.87	0.88	0.84	0.78	0.7	0.69	0.7
$d_{M3\text{-}C}$				-0.06				0.05				-0.09
Agreement Indices												
	M1–M2	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**
Exact	55%	62%	55%	62%	60%	58%	57%	59%	63%	64%	66%	67%
Adjacent	44%	37%	42%	37%	38%	40%	41%	38%	35%	35%	32%	32%
Non-Adjacent	2%	2%	3%	0.70%	3%	2%	3%	3%	2%	0.90%	2%	0.90%
Kappa	0.33	0.43	0.33	0.44	0.4	0.36	0.36	0.38	0.37	0.38	0.41	0.42
QWK	0.67	0.72	0.65	0.74	0.69	0.68	0.68	0.66	0.61	0.64	0.63	0.63
Pearson *r*	0.68	0.72	0.66	0.74	0.7	0.68	0.68	0.66	0.61	0.65	0.63	0.63

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 6. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for N2_579

	Paragraphing				Sentence Structure				Punctuation			
	M1	M2	M3	C	M1	M2	M3	C	M1	M2	M3	C
N	435	433	434	434	435	433	434	434	435	433	434	434
Score Distributions												
0	32%	28%	30%	28%	2%	1%	1%	1%	3%	2%	2%	2%
1	45%	55%	51%	55%	4%	5%	5%	3%	14%	12%	16%	14%
2	22%	16%	19%	17%	19%	20%	13%	19%	23%	30%	30%	32%
3					41%	43%	41%	37%	42%	41%	38%	38%
4					26%	25%	33%	36%	16%	13%	13%	13%
5					8%	6%	6%	4%	3%	1%	0.70%	0.90%
6					1%	**0.00%**	**0.00%**	**0.00%**				
Mean	0.9	0.88	0.89	0.9	3.13	3.04	3.18	3.16	2.64	2.54	2.46	2.5
SD	0.73	0.66	0.69	0.66	1.07	0.99	1.01	0.97	1.08	0.99	1	0.98
d_{M3-C}				-0.01				0.02				-0.04
Agreement Indices												
	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3
Exact	68%	65%	69%	65%	50%	51%	50%	54%	54%	51%	58%	55%
Adjacent	32%	34%	30%	34%	45%	45%	45%	43%	42%	46%	38%	41%
Non-Adjacent	0.50%	2%	0.90%	1%	5%	4%	5%	4%	4%	3%	4%	3%
Kappa	0.48	0.43	0.51	0.42	0.3	0.32	0.3	0.34	0.36	0.32	0.42	0.38
QWK	0.65	0.59	0.67	0.58	0.66	0.7	0.69	0.7	0.73	0.71	0.75	0.71
Pearson *r*	0.66	0.59	0.67	0.58	0.67	0.7	0.69	0.7	0.73	0.72	0.76	0.71

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 7. Marker and CRASE Agreement for Spelling for N2_579

	Spelling			
	M1	M2	M3	C
N	435	433	434	434
Score Distributions				
0	1%	0.70%	0.70%	1%
1	3%	2%	2%	1%
2	8%	10%	8%	6%
3	34%	33%	28%	30%
4	40%	34%	38%	37%
5	12%	18%	23%	23%
6	2%	2%	2%	1%
Mean	3.53	3.59	3.74	3.74
SD	1.03	1.06	1.05	1.03
d_{M3-C}				0
Agreement Indices				
	M1–M2	M1–C	M1–M3	C–M3
Exact	49%	47%	47%	52%
Adjacent	45%	48%	48%	43%
Non-Adjacent	5%	5%	6%	5%
Kappa	0.29	0.26	0.26	0.33
QWK	0.68	0.69	0.68	0.71
Pearson *r*	0.68	0.7	0.69	0.71

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Prompt P2_579

Table 8. Marker and CRASE agreement for audience, text structure and ideas for P2_579

	Audience				Text Structure				Ideas			
	M1	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**
N	518	518	517	518	518	518	517	518	518	518	517	518
Score Distributions												
0	1%	1%	**0.00%**	0.60%	2%	2%	2%	2%	0.80%	0.80%	0.80%	1%
1	2%	3%	3%	3%	13%	17%	17%	15%	5%	6%	3%	4%
2	15%	19%	15%	14%	43%	39%	37%	42%	22%	21%	15%	15%
3	39%	37%	37%	42%	32%	34%	35%	35%	47%	49%	54%	54%
4	29%	27%	28%	24%	10%	8%	9%	6%	22%	19%	22%	23%
5	11%	11%	12%	16%					3%	4%	6%	2%
6	3%	3%	4%	1%								
Mean	3.37	3.3	3.44	3.37	2.34	2.29	2.33	2.28	2.95	2.92	3.11	3
SD	1.09	1.12	1.11	1.08	0.9	0.91	0.92	0.87	0.92	0.92	0.89	0.87
$d_{M3\text{-}C}$				0.06				0.06				0.12
Agreement Indices												
	M1–M2	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**
Exact	50%	55%	53%	60%	56%	61%	58%	61%	54%	59%	57%	67%
Adjacent	45%	41%	43%	37%	41%	38%	41%	37%	42%	38%	39%	32%
Non-Adjacent	5%	3%	4%	3%	3%	1%	1%	2%	3%	3%	3%	1%
Kappa	0.32	0.39	0.35	0.45	0.37	0.43	0.4	0.44	0.33	0.38	0.36	0.48
QWK	0.7	0.75	0.75	0.79	0.68	0.7	0.72	0.71	0.67	0.69	0.69	0.76
Pearson *r*	0.71	0.75	0.75	0.8	0.68	0.7	0.72	0.72	0.67	0.69	0.7	0.77

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 9. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P2_579

	Persuasive Devices				Vocabulary				Cohesion			
	M1	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**	**M1**	**M2**	**M3**	**C**
N	518	518	517	518	518	518	517	518	518	518	517	518
Score Distributions												
0	3%	3%	1%	2%	0.80%	0.60%	0.40%	0.80%	1%	1%	1%	0.80%
1	19%	23%	15%	15%	3%	3%	2%	2%	7%	7%	5%	7%
2	45%	39%	39%	43%	36%	37%	39%	42%	55%	57%	55%	59%
3	26%	27%	33%	31%	41%	43%	40%	35%	32%	30%	34%	32%
4	8%	7%	11%	9%	16%	14%	14%	19%	5%	5%	5%	2%
5					3%	3%	4%	1%				
6												
Mean	2.17	2.12	2.38	2.29	2.79	2.75	2.77	2.73	2.33	2.31	2.36	2.26
SD	0.91	0.95	0.92	0.91	0.89	0.85	0.88	0.86	0.72	0.71	0.7	0.64
$d_{M3\text{-}C}$				0.1				0.05				**0.15**
Agreement Indices												
	M1–M2	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**	**M1–M2**	**M1–C**	**M1–M3**	**C–M3**
Exact	49%	58%	54%	57%	59%	61%	59%	64%	64%	68%	65%	69%
Adjacent	43%	36%	40%	42%	39%	37%	39%	34%	34%	31%	34%	30%
Non-Adjacent	8%	5%	5%	2%	2%	2%	2%	1%	3%	1%	0.80%	0.60%
Kappa	0.27	0.39	0.35	0.38	0.38	0.42	0.38	0.47	0.37	0.44	0.4	0.45
QWK	0.57	0.63	0.61	0.71	0.68	0.7	0.7	0.74	0.56	0.62	0.63	0.64
Pearson *r*	0.57	0.64	0.62	0.71	0.68	0.7	0.7	0.74	0.56	0.62	0.63	0.65

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 10. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for P2_579

	Paragraphing				Sentence Structure				Punctuation			
	M1	M2	M3	C	M1	M2	M3	C	M1	M2	M3	C
N	518	518	517	518	518	518	517	518	518	518	517	518
Score Distributions												
0	17%	18%	16%	15%	1%	1%	1%	1%	3%	3%	2%	2%
1	31%	36%	38%	42%	4%	3%	4%	3%	11%	9%	12%	14%
2	38%	35%	37%	34%	18%	19%	17%	19%	23%	25%	25%	29%
3	13%	12%	9%	9%	41%	47%	43%	43%	45%	44%	43%	40%
4					29%	23%	29%	29%	16%	16%	17%	15%
5					6%	5%	7%	6%	2%	3%	0.80%	0.80%
6					1%	2%	0.40%	**0.00%**				
Mean	1.48	1.41	1.39	1.38	3.16	3.09	3.16	3.12	2.69	2.71	2.64	2.56
SD	0.93	0.91	0.85	0.85	1.03	0.98	0.98	0.96	1.03	1.03	1	0.99
d$_{M3-C}$				0.01				0.04				0.08
Agreement Indices												
	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3	M1–M2	M1–C	M1–M3	C–M3
Exact	57%	59%	65%	64%	53%	55%	51%	56%	55%	54%	59%	**53%**
Adjacent	41%	39%	32%	34%	42%	41%	44%	41%	41%	41%	38%	42%
Non-Adjacent	2%	2%	2%	1%	5%	4%	4%	3%	4%	5%	3%	5%
Kappa	0.4	0.42	0.5	0.48	0.33	0.36	0.31	0.36	0.37	0.36	0.42	0.34
QWK	0.71	0.7	0.74	0.72	0.67	0.69	0.69	0.7	0.73	0.7	0.76	0.69
Pearson *r*	0.71	0.7	0.74	0.72	0.67	0.69	0.69	0.7	0.73	0.7	0.77	0.69

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.

Table 11. Marker and CRASE agreement for spelling for P2_579

	Spelling			
	M1	M2	M3	C
N	518	518	517	518
Score Distributions				
0	0.80%	1%	0.60%	0.80%
1	2%	2%	2%	2%
2	6%	7%	7%	8%
3	36%	33%	22%	22%
4	40%	38%	41%	39%
5	14%	17%	25%	26%
6	1%	2%	2%	2%
Mean	3.6	3.63	3.85	3.83
SD	0.96	1.03	1.04	1.06
d$_{M3-C}$				0.02
Agreement Indices				
	M1–M2	M1–C	M1–M3	C–M3
Exact	54%	54%	51%	58%
Adjacent	44%	44%	46%	39%
Non-Adjacent	2%	3%	3%	2%
Kappa	0.34	0.36	0.31	0.42
QWK	0.72	0.73	0.71	0.77
Pearson *r*	0.72	0.75	0.74	0.77

Note. M1=marker 1; M2=marker 2; M3=expert marker; C=CRASE.intra-criteria correlations

Prompt P1_357

Table 12. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for P1_357

Source	N	Mean	St. D.	Min	Max
Marker 1	614	0.67	0.07	0.53	0.79
Marker 2	614	0.68	0.06	0.55	0.81
Marker 3	615	0.68	0.07	0.53	0.81
CRASE	615	0.79	0.07	0.64	0.93

Table 13. Intra-criteria correlations of M1 (top), M2 (bottom) for P1_357

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.79	0.77	0.75	0.74	0.7	0.71	0.77	0.57	0.7
TS	0.8		0.71	0.79	0.68	0.67	0.72	0.71	0.58	0.66
ID	0.81	0.75		0.69	0.7	0.68	0.63	0.71	0.58	0.69
PD	0.76	0.81	0.66		0.65	0.63	0.67	0.68	0.54	0.63
VO	0.74	0.68	0.69	0.64		0.68	0.63	0.75	0.55	0.71
CO	0.73	0.69	0.69	0.66	0.67		0.63	0.71	0.55	0.64
PA	0.72	0.71	0.68	0.64	0.61	0.61		0.64	0.53	0.6
SS	0.74	0.71	0.72	0.66	0.7	0.72	0.63		0.66	0.72
PU	0.59	0.57	0.6	0.55	0.56	0.6	0.58	0.66		0.65
SP	0.72	0.69	0.71	0.66	0.66	0.67	0.6	0.73	0.6	

Table 14. Intra-criteria correlations of C (top), M3 (bottom) for P1_357

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.89	0.85	0.88	0.77	0.74	0.82	0.92	0.81	0.84
TS	0.79		0.8	0.93	0.73	0.71	0.89	0.83	0.79	0.82
ID	0.81	0.72		0.8	0.73	0.73	0.71	0.87	0.8	0.83
PD	0.79	0.81	0.72		0.71	0.73	0.84	0.85	0.78	0.82
VO	0.77	0.66	0.7	0.68		0.67	0.69	0.75	0.7	0.76
CO	0.69	0.61	0.67	0.63	0.69		0.64	0.77	0.73	0.75
PA	0.76	0.76	0.66	0.71	0.63	0.57		0.76	0.74	0.8
SS	0.79	0.68	0.72	0.71	0.7	0.68	0.67		0.83	0.85
PU	0.62	0.58	0.57	0.55	0.53	0.54	0.59	0.66		0.85
SP	0.77	0.7	0.69	0.69	0.69	0.65	0.69	0.75	0.64	

Prompt N1_357

Table 15. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for N1_357

Source	N	Mean	St. D.	Min	Max
Marker 1	546	0.64	0.1	0.48	0.81
Marker 2	546	0.61	0.1	0.43	0.77
Marker 3	545	0.64	0.1	0.45	0.85
CRASE	547	0.77	0.07	0.53	0.89

Table 16. Intra-criteria correlations of M1 (top), M2 (bottom) for N1_357

	AU	TS	ID	CS	VO	CO	PA	SS	PU	SP
AU		0.76	0.81	0.77	0.78	0.68	0.55	0.75	0.55	0.68
TS	0.73		0.74	0.75	0.66	0.67	0.55	0.65	0.49	0.59
ID	0.73	0.71		0.76	0.76	0.69	0.52	0.69	0.55	0.66
CS	0.73	0.77	0.72		0.71	0.67	0.49	0.67	0.5	0.61
VO	0.73	0.64	0.7	0.7		0.69	0.5	0.76	0.53	0.69
CO	0.62	0.62	0.68	0.65	0.68		0.49	0.69	0.54	0.64
PA	0.5	0.47	0.5	0.47	0.45	0.48		0.54	0.5	0.48
SS	0.64	0.59	0.67	0.62	0.72	0.7	0.47		0.64	0.73
PU	0.5	0.48	0.54	0.51	0.54	0.56	0.43	0.66		0.63
SP	0.63	0.57	0.64	0.63	0.69	0.66	0.45	0.72	0.61	

Table 17. Intra-criteria correlations of C (top), M3 (bottom) for N1_357

	AU	TS	ID	CS	VO	CO	PA	SS	PU	SP
AU		0.82	0.86	0.82	0.8	0.74	0.75	0.89	0.8	0.83
TS	0.72		0.78	0.89	0.75	0.69	0.73	0.82	0.77	0.8
ID	0.85	0.69		0.79	0.75	0.76	0.71	0.88	0.75	0.83
CS	0.74	0.79	0.73		0.8	0.72	0.71	0.83	0.79	0.85
VO	0.8	0.62	0.75	0.71		0.74	0.56	0.78	0.74	0.82
CO	0.72	0.62	0.72	0.67	0.72		0.53	0.81	0.71	0.75
PA	0.57	0.49	0.56	0.51	0.5	0.51		0.75	0.7	0.72
SS	0.77	0.59	0.75	0.66	0.73	0.71	0.53		0.84	0.85
PU	0.59	0.49	0.58	0.51	0.56	0.52	0.45	0.67		0.84
SP	0.72	0.58	0.71	0.67	0.68	0.63	0.5	0.74	0.64	

Prompt N2_579

Table 18. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for N2_579

Source	N	Mean	St. D.	Min	Max
Marker 1	435	0.7	0.09	0.46	0.82
Marker 2	433	0.67	0.1	0.49	0.85
Marker 3	434	0.7	0.1	0.48	0.88
CRASE	435	0.83	0.06	0.65	0.92

Table 19. Intra-criteria correlations of M1 (top), M2 (bottom) for Prompt N2_579

	AU	TS	ID	CS	VO	CO	PA	SS	PU	SP
AU		0.82	0.81	0.8	0.82	0.77	0.58	0.78	0.59	0.64
TS	0.81		0.82	0.8	0.76	0.76	0.58	0.73	0.6	0.61
ID	0.85	0.78		0.81	0.8	0.77	0.62	0.77	0.66	0.67
CS	0.84	0.76	0.79		0.79	0.79	0.62	0.76	0.63	0.68
VO	0.8	0.69	0.74	0.74		0.8	0.61	0.8	0.64	0.72
CO	0.76	0.68	0.69	0.73	0.73		0.6	0.8	0.65	0.66
PA	0.65	0.62	0.6	0.6	0.57	0.54		0.59	0.53	0.46
SS	0.77	0.67	0.73	0.72	0.72	0.71	0.54		0.72	0.72
PU	0.57	0.51	0.57	0.52	0.53	0.54	0.49	0.64		0.65
SP	0.74	0.63	0.68	0.71	0.73	0.67	0.49	0.71	0.57	

Table 20. Intra-criteria correlations of C (top), M3 (bottom) for N2_579

	AU	TS	ID	CS	VO	CO	PA	SS	PU	SP
AU		0.91	0.9	0.91	0.9	0.87	0.79	0.92	0.84	0.88
TS	0.83		0.84	0.92	0.85	0.83	0.77	0.9	0.79	0.84
ID	0.88	0.81		0.84	0.83	0.8	0.78	0.85	0.79	0.87
CS	0.86	0.85	0.82		0.89	0.84	0.74	0.86	0.81	0.87
VO	0.85	0.74	0.8	0.79		0.82	0.71	0.83	0.78	0.86
CO	0.77	0.74	0.74	0.77	0.75		0.65	0.87	0.78	0.84
PA	0.62	0.6	0.6	0.58	0.59	0.51		0.75	0.69	0.73
SS	0.8	0.7	0.76	0.76	0.76	0.76	0.58		0.85	0.88
PU	0.64	0.54	0.59	0.58	0.61	0.59	0.48	0.69		0.85
SP	0.76	0.66	0.7	0.71	0.75	0.7	0.51	0.76	0.68	

Prompt P2_579

Table 21. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for P2_579

Source	N	Mean	St. D.	Min	Max
Marker 1	518	0.71	0.06	0.55	0.84
Marker 2	518	0.7	0.07	0.54	0.85
Marker 3	517	0.74	0.08	0.56	0.88
CRASE	518	0.86	0.04	0.79	0.94

Table 22. Intra-criteria correlations of M1 (top), M2 (bottom) for P2_579

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.82	0.84	0.81	0.79	0.74	0.73	0.77	0.64	0.76
TS	0.82		0.75	0.73	0.72	0.73	0.76	0.76	0.62	0.68
ID	0.85	0.78		0.73	0.79	0.75	0.72	0.75	0.63	0.74
PD	0.82	0.76	0.74		0.7	0.68	0.64	0.65	0.55	0.69
VO	0.81	0.71	0.79	0.72		0.73	0.67	0.77	0.62	0.73
CO	0.74	0.71	0.74	0.66	0.73		0.67	0.74	0.62	0.66
PA	0.74	0.74	0.71	0.67	0.68	0.67		0.68	0.59	0.63
SS	0.75	0.71	0.75	0.64	0.78	0.77	0.68		0.74	0.73
PU	0.61	0.6	0.63	0.54	0.62	0.6	0.56	0.71		0.69
SP	0.72	0.63	0.7	0.65	0.7	0.63	0.58	0.69	0.65	

Table 23. Intra-criteria correlations of C (top), M3 (bottom) for P2_579

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.92	0.9	0.94	0.89	0.87	0.89	0.9	0.87	0.88
TS	0.86		0.87	0.94	0.84	0.84	0.91	0.91	0.85	0.86
ID	0.88	0.81		0.87	0.82	0.85	0.82	0.89	0.83	0.88
PD	0.85	0.84	0.81		0.84	0.85	0.9	0.92	0.84	0.86
VO	0.85	0.75	0.82	0.8		0.82	0.82	0.84	0.83	0.84
CO	0.81	0.76	0.8	0.76	0.78		0.79	0.86	0.79	0.84
PA	0.78	0.79	0.73	0.75	0.71	0.71		0.85	0.83	0.82
SS	0.81	0.76	0.8	0.76	0.74	0.77	0.72		0.86	0.89
PU	0.63	0.59	0.66	0.56	0.57	0.6	0.6	0.69		0.88
SP	0.76	0.69	0.77	0.71	0.75	0.68	0.63	0.74	0.66	

Summed scores correlations

Table 24. Summed score distribution statistics for P1_357

	M1	M2	M3	CRASE
N	615	615	615	615
Mean	20.46	20.19	20.71	20.27
SD	7.01	6.99	6.82	7
Min	0	0	0	0
Max	45	43	44	36
	M1–M2	**C-M1**	**M1–M3**	**C–M3**
Correlation	0.85	0.84	0.86	0.85

Table 35. Summed score distribution statistics for N1_357

	M1	M2	M3	CRASE
N	547	547	547	547
Mean	20.96	20.36	21.37	21.37
SD	6.93	6.57	6.72	6.92
Min	0	0	0	0
Max	46	46	41	44
	M1–M2	**C-M1**	**M1–M3**	**C–M3**
Correlation	0.87	0.84	0.88	0.85

Table 25. Summed score distribution statistics for N2_579

	M1	M2	M3	CRASE
N	435	435	435	435
Mean	25.84	25.21	26.23	26.34
SD	8.32	7.88	7.97	8.41
Min	0	0	0	0
Max	47	45	46	46
	M1–M2	**C-M1**	**M1–M3**	**C–M3**
Correlation	0.86	0.84	0.85	0.83

Table 37. Summed score distribution statistics for P2_579

	M1	M2	M3	CRASE
N	518	518	517	518
Mean	26.89	26.51	27.38	26.83
SD	8.07	8.06	8.21	8.53
Min	0	0	0	0
Max	48	48	47	47
	M1–M2	**C-M1**	**M1–M3**	**C–M3**
Correlation	0.85	0.84	0.86	0.85

Appendix 1b. Replication analyses blind validation sample results

Criteria scores correlations

The criteria-level results appearing in this section are in relation to the blind validation sample for each writing prompt. The results in the table are obtained from the models built only upon the training sample. The cuts applied to the CRASE continuous scores were set to match the distribution of marker 1.

Table cells shaded in orange represent score points for which there were no records obtaining the score from marker 1, marker 2, or CRASE. Table cells shaded in light red represent CRASE metrics that did not meet the established thresholds.

Standards for a criteria passing a threshold were:

- The absolute standardised mean difference (SMD) between CRASE and marker 1 is 0.15 or lower.
- The CRASE – marker 1 exact agreement rate is 5 per cent or lower than the marker 1 – marker 2 exact agreement rate.
- The quadratic weighted kappa (QWK) statistic for CRASE and marker 1 is 0.10 or lower than the marker 1 – marker 2 QWK.

Prompt P3_357

Table 1. Marker and CRASE agreement for audience, text structure and ideas for P3_357

	Audience			Text Structure			Ideas		
	M1	M2	C	M1	M2	C	M1	M2	C
N	142	142	142	142	142	142	142	142	142
Score Distributions									
0	0 %	0.7%	0.7%	0.7%	1%	3%	0.7%	0.7%	0.7%
1	5%	4%	4%	56%	51%	51%	5%	4%	6%
2	42%	37%	35%	32%	34%	37%	43%	40%	41%
3	39%	42%	46%	11%	11%	9%	46%	46%	49%
4	13%	14%	13%	0 %	2%	0 %	5%	8%	4%
5	2%	3%	1%				0.7%	0.7%	0 %
6	0 %	0 %	0 %						
Mean	2.65	2.73	2.70	1.54	1.61	1.53	2.51	2.61	2.50
SD	0.84	0.88	0.83	0.70	0.79	0.70	0.73	0.75	0.70
$d_{M3\text{-}C}$			0.06			-0.01			-0.01
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	70%	75%	68%	68%	71%	65%	68%	70%	74%
Adjacent	30%	25%	31%	32%	28%	35%	32%	30%	25%
Non-Adjacent	0.7%	0%	0.7%	0.7%	0.7%	0%	0%	0%	0.7%
Kappa	0.55	0.62	0.52	0.45	0.52	0.41	0.48	0.50	0.56
QWK	0.78	0.83	0.76	0.69	0.72	0.65	0.71	0.72	0.72
Pearson *r*	0.79	0.83	0.76	0.70	0.73	0.65	0.72	0.72	0.72

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 2. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P3_357

	Persuasive Devices			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	142	142	142	142	142	142	142	142	142
Score Distributions									
0	0.7%	1%	0.7%	0.7%	0.7%	0.7%	0.7%	2%	0.7%
1	24%	30%	26%	6%	4%	5%	13%	13%	13%
2	56%	49%	57%	56%	61%	59%	67%	65%	73%
3	19%	18%	15%	32%	30%	31%	18%	18%	13%
4	0.7%	2%	0.7%	5%	5%	4%	0.70%	1%	0 %
5				0 %	0 %	0 %			
6									
Mean	1.95	1.89	1.89	2.35	2.34	2.33	2.05	2.04	1.99
SD	0.70	0.78	0.68	0.71	0.67	0.67	0.61	0.67	0.54
$d_{M3\text{-}C}$			-0.09			-0.03			-0.10
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	69%	70%	70%	68%	74%	69%	72%	76%	75%
Adjacent	30%	30%	30%	32%	26%	31%	27%	23%	24%
Non-Adjacent	1%	0%	0.7%	0.7%	0%	0%	0.7%	0.7%	0.7%
Kappa	0.50	0.52	0.49	0.42	0.52	0.45	0.45	0.50	0.47
QWK	0.68	0.72	0.66	0.63	0.71	0.67	0.63	0.65	0.60
Pearson *r*	0.68	0.73	0.66	0.63	0.71	0.67	0.63	0.67	0.60

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 3. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for P3_357

	Paragraphing			Sentence Structure			Punctuation		
	M1	**M2**	**C**	**M1**	**M2**	**C**	**M1**	**M2**	**C**
N	142	142	142	142	142	142	142	142	142
Score Distributions									
0	34%	35%	37%	1%	1%	0.7%	6%	7%	6%
1	34%	39%	35%	7%	8%	7%	18%	20%	18%
2	32%	25%	27%	42%	37%	37%	41%	45%	41%
3	0.7%	0.7%	0.7%	35%	32%	37%	35%	26%	32%
4				15%	21%	18%	**0%**	2%	1%
5				**0%**	**0%**	0.7%	**0%**	**0%**	1%
6				**0%**	**0%**	**0%**			
Mean	0.99	0.91	0.92	2.54	2.63	2.67	2.04	1.96	2.10
SD	0.83	0.79	0.82	0.88	0.96	0.91	0.89	0.91	0.96
d$_{M3\text{-}C}$			-0.08			0.15			0.06
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	79%	74%	**73%**	61%	56%	57%	72%	64%	**65%**
Adjacent	21%	25%	25%	39%	42%	42%	28%	35%	33%
Non-Adjacent	0%	0.7%	1%	0%	2%	0.7%	0%	1%	2%
Kappa	0.68	0.61	0.6	0.44	0.38	0.37	0.59	0.48	0.49
QWK	0.84	0.78	0.77	0.77	0.71	0.72	0.82	0.77	0.75
Pearson *r*	0.84	0.78	0.77	0.78	0.71	0.73	0.83	0.78	0.76

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 4. Marker and CRASE agreement for spelling for P3_357

	Spelling		
	M1	**M2**	**C**
N	142	142	142
Score Distributions			
0	0.7%	0.7%	0.7%
1	4%	6%	2%
2	18%	23%	23%
3	39%	37%	37%
4	32%	27%	30%
5	6%	7%	8%
6	**0%**	**0%**	**0%**
Mean	3.18	3.06	3.16
SD	0.97	1.04	0.98
$d_{M3\text{-}C}$			-0.02
Agreement Indices			
	M1–M2	**C-M2**	**C–M1**
Exact	71%	63%	**58%**
Adjacent	29%	35%	42%
Non-Adjacent	0%	1%	0.7%
Kappa	0.60	0.50	0.41
QWK	0.86	0.80	0.77
Pearson *r*	0.86	0.80	0.77

Note. M1=marker 1; M2=marker 2; C=CRASE.

Prompt P4_357

Table 5. Marker and CRASE agreement for audience, text structure and ideas for P4_357

	Audience			Text Structure			Ideas		
	M1	M2	C	M1	M2	C	M1	M2	C
N	125	125	125	125	125	125	125	125	125
Score Distributions									
0	**0%**	**0%**	**0%**	2%	4%	0.8%	**0%**	0.8%	**0%**
1	3%	7%	2%	50%	42%	47%	5%	6%	5%
2	46%	42%	44%	34%	36%	36%	50%	47%	46%
3	33%	32%	39%	11%	15%	14%	37%	36%	38%
4	14%	15%	9%	2%	2%	2%	8%	9%	11%
5	3%	3%	6%				0.8%	0.8%	**0%**
6	0.8%	0.8%	**0%**						
Mean	2.70	2.68	2.74	1.61	1.70	1.69	2.50	2.48	2.55
SD	0.92	0.98	0.89	0.81	0.86	0.79	0.75	0.81	0.76
d_{M3-C}			0.04			0.10			0.07
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	68%	67%	74%	66%	70%	69%	78%	82%	78%
Adjacent	32%	32%	26%	34%	30%	30%	22%	18%	22%
Non-Adjacent	0%	0.8%	0%	0%	0%	0.8%	0.8%	0%	0%
Kappa	0.53	0.51	0.61	0.47	0.54	0.50	0.64	0.71	0.64
QWK	0.82	0.80	0.84	0.75	0.78	0.74	0.79	0.85	0.80
Pearson *r*	0.82	0.80	0.84	0.76	0.78	0.74	0.80	0.86	0.80

Table 6. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P4_357

	Persuasive Devices			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	125	125	125	125	125	125	125	125	125
Score Distributions									
0	2%	4%	2%	**0%**	**0%**	**0%**	**0%**	**0%**	0.8%
1	22%	27%	21%	3%	6%	3%	14%	14%	6%
2	58%	49%	61%	72%	68%	67%	68%	67%	76%
3	13%	18%	12%	17%	17%	23%	14%	17%	12%
4	4%	2%	5%	6%	8%	5%	3%	2%	5%
5				2%	0.8%	2%			
6									
Mean	1.94	1.86	1.98	2.31	2.29	2.34	2.06	2.06	2.14
SD	0.78	0.82	0.77	0.71	0.74	0.70	0.64	0.61	0.63
d_{M3-C}			0.05			0.04			0.13
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	66%	63%	69%	77%	70%	74%	82%	77%	80%
Adjacent	33%	34%	30%	23%	28%	26%	18%	22%	20%
Non-Adjacent	0.8%	3%	2%	0%	2%	0.8%	0%	0.8%	0%
Kappa	0.47	0.41	0.46	0.51	0.41	0.44	0.63	0.49	0.56
QWK	0.72	0.63	0.70	0.78	0.66	0.71	0.77	0.67	0.75
Pearson *r*	0.72	0.64	0.70	0.78	0.67	0.71	0.77	0.67	0.76

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 7. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for P4_357

	Paragraphing			Sentence Structure			Punctuation		
	M1	**M2**	**C**	**M1**	**M2**	**C**	**M1**	**M2**	**C**
N	125	125	125	125	125	125	125	125	125
Score Distributions									
0	33%	35%	34%	**0 %**	0.8%	**0 %**	2%	0.8%	2%
1	47%	43%	46%	9%	10%	6%	20%	22%	19%
2	14%	19%	18%	46%	46%	40%	38%	36%	38%
3	6%	2%	2%	26%	26%	37%	30%	31%	34%
4				15%	14%	14%	10%	9%	6%
5				3%	3%	2%	0.8%	0.8%	2%
6				0.8%	**0 %**	0.8%			
Mean	0.93	0.89	0.88	2.61	2.52	2.70	2.28	2.27	2.29
SD	0.83	0.80	0.78	1.01	0.98	0.93	0.97	0.96	0.96
$d_{M3\text{-}C}$			-0.06			0.09			0.01
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	77%	74%	73%	64%	50%	59%	70%	66%	74%
Adjacent	23%	26%	27%	34%	46%	38%	28%	34%	25%
Non-Adjacent	0%	0%	0%	2%	5%	2%	2%	0.8%	2%
Kappa	0.64	0.59	0.58	0.48	0.27	0.41	0.59	0.52	0.63
QWK	0.82	0.79	0.79	0.79	0.65	0.74	0.81	0.80	0.83
Pearson *r*	0.83	0.79	0.79	0.80	0.66	0.75	0.81	0.80	0.83

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 8. Marker and CRASE agreement for spelling for P4_357

	Spelling		
	M1	M2	C
N	125	125	125
Score Distributions			
0	0 %	0 %	0 %
1	6%	6%	2%
2	25%	24%	26%
3	31%	30%	42%
4	25%	26%	14%
5	13%	11%	12%
6	0.8%	2%	2%
Mean	3.17	3.21	3.14
SD	1.13	1.17	1.09
d_{M3-C}			-0.03
Agreement Indices			
	M1–M2	C-M2	C–M1
Exact	74%	59%	72%
Adjacent	22%	40%	28%
Non-Adjacent	4%	0.8%	0%
Kappa	0.66	0.46	0.63
QWK	0.86	0.83	0.89
Pearson *r*	0.86	0.83	0.89

Note. M1=marker 1; M2=marker 2; C=CRASE.

Prompt P5_357

Table 9. Marker and CRASE agreement for audience, text structure and ideas for P5_357

	Audience			Text Structure			Ideas		
	M1	M2	C	M1	M2	C	M1	M2	C
N	128	128	128	128	128	128	128	128	128
Score Distributions									
0	0.80%	0.80%	0%	2%	3%	2%	0.8%	0.8%	0%
1	3%	5%	8%	53%	52%	49%	5%	8%	9%
2	50%	42%	44%	30%	34%	38%	59%	51%	54%
3	30%	38%	34%	15%	9%	11%	25%	36%	37%
4	13%	13%	14%	0%	0.80%	0%	9%	5%	0.8%
5	3%	2%	0%				0.80%	0%	0%
6	0%	0%	0%						
Mean	2.60	2.63	2.55	1.59	1.52	1.59	2.39	2.36	2.30
SD	0.89	0.86	0.83	0.76	0.74	0.70	0.80	0.73	0.63
$d_{M3\text{-}C}$			-0.06			0.00			-0.12
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	75%	66%	63%	70%	64%	72%	72%	73%	71%
Adjacent	23%	33%	37%	30%	36%	28%	28%	27%	28%
Non-Adjacent	2%	0.8%	0%	0%	0%	0%	0%	0%	0.8%
Kappa	0.62	0.49	0.44	0.51	0.40	0.54	0.53	0.55	0.51
QWK	0.79	0.75	0.75	0.74	0.66	0.74	0.76	0.71	0.70
Pearson *r*	0.79	0.75	0.75	0.74	0.66	0.74	0.76	0.72	0.72

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table10. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P5_357

	Persuasive Devices			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	128	128	128	128	128	128	128	128	128
Score Distributions									
0	2%	2%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%	0.8%
1	23%	24%	26%	4%	3%	3%	13%	13%	11%
2	56%	50%	56%	70%	70%	68%	68%	66%	71%
3	18%	23%	16%	20%	20%	27%	19%	20%	17%
4	0.8%	0%	0.8%	5%	5%	0.8%	0%	0%	0%
5				0.8%	0%	0%			
6									
Mean	1.93	1.95	1.91	2.27	2.27	2.24	2.05	2.06	2.05
SD	0.71	0.76	0.69	0.70	0.65	0.56	0.59	0.60	0.56
$d_{M3\text{-}C}$			-0.03			-0.05			0.00
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	69%	70%	76%	79%	82%	79%	77%	83%	84%
Adjacent	31%	30%	24%	20%	18%	20%	23%	17%	16%
Non-Adjacent	0%	0%	0%	2%	0%	0.8%	0%	0%	0%
Kappa	0.50	0.51	0.59	0.55	0.61	0.55	0.53	0.64	0.67
QWK	0.71	0.71	0.75	0.71	0.75	0.70	0.66	0.74	0.76
Pearson *r*	0.71	0.71	0.75	0.71	0.76	0.72	0.66	0.74	0.76

Note. M1=marker 1; M2=marker 2; C=CRASE

Table 11. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for P5_357

	Paragraphing			Sentence Structure			Punctuation		
	M1	**M2**	**C**	**M1**	**M2**	**C**	**M1**	**M2**	**C**
N	128	128	128	128	128	128	128	128	128
Score Distributions									
0	42%	44%	45%	0.8%	0.8%	0.8%	3%	2%	3%
1	35%	35%	35%	5%	9%	8%	18%	14%	19%
2	20%	20%	20%	49%	40%	42%	42%	44%	45%
3	2%	2%	**0%**	29%	39%	36%	28%	36%	24%
4				15%	11%	13%	7%	4%	9%
5				2%	0.8%	**0%**	2%	**0%**	**0%**
6				**0%**	**0%**	**0%**			
Mean	0.83	0.79	0.74	2.57	2.53	2.53	2.23	2.25	2.16
SD	0.83	0.81	0.77	0.88	0.86	0.85	0.98	0.83	0.94
$d_{M3\text{-}C}$			-0.11			-0.05			-0.07
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	68%	68%	73%	60%	63%	63%	66%	60%	**50%**
Adjacent	32%	31%	27%	38%	38%	37%	32%	38%	48%
Non-Adjacent	0%	0.8%	0%	2%	0%	0%	2%	2%	2%
Kappa	0.51	0.50	0.58	0.41	0.44	0.45	0.50	0.42	0.29
QWK	0.76	0.72	0.79	0.71	0.74	0.75	0.75	0.72	0.70
Pearson *r*	0.76	0.72	0.79	0.71	0.74	0.76	0.76	0.72	0.70

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 12. Marker and CRASE agreement for spelling for P5_357

	Spelling		
	M1	**M2**	**C**
N	128	128	128
Score Distributions			
0	0.8%	0.8%	2%
1	2%	5%	2%
2	27%	27%	31%
3	33%	30%	30%
4	27%	27%	28%
5	10%	11%	7%
6	**0 %**	0.8%	**0 %**
Mean	3.14	3.13	3.02
SD	1.05	1.14	1.06
d_{M3-C}			-0.11
Agreement Indices			
	M1–M2	**C-M2**	**C–M1**
Exact	64%	58%	64%
Adjacent	36%	41%	34%
Non-Adjacent	0%	2%	2%
Kappa	0.52	0.43	0.51
QWK	0.85	0.80	0.82
Pearson *r*	0.85	0.81	0.82

Note. M1=marker 1; M2=marker 2; C=CRASE.

Prompt N3_357

Table 13. Marker and CRASE agreement for audience, text structure and ideas for N3_357

	Audience			Text Structure			Ideas		
	M1	M2	C	M1	M2	C	M1	M2	C
N	154	154	154	154	154	154	154	154	154
Score Distributions									
0	0 %	0 %	0 %	0.7%	1%	0.7%	0.7%	1%	0.7%
1	1%	3%	4%	31%	31%	32%	12%	6%	6%
2	45%	44%	49%	55%	51%	53%	44%	45%	58%
3	42%	37%	32%	13%	16%	15%	40%	43%	31%
4	10%	15%	12%	0.7%	0.7%	0 %	4%	3%	4%
5	2%	2%	2%				0 %	0.7%	0 %
6	0 %	0 %	0 %						
Mean	2.67	2.70	2.59	1.82	1.84	1.82	2.34	2.42	2.31
SD	0.76	0.83	0.83	0.68	0.73	0.68	0.76	0.75	0.68
$d_{M3\text{-}C}$			-0.10			0.00			-0.04
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	68%	68%	70%	64%	65%	73%	69%	74%	71%
Adjacent	31%	32%	29%	34%	34%	27%	30%	25%	28%
Non-Adjacent	2%	0%	1%	1%	0.7%	0%	0.7%	0.7%	0.7%
Kappa	0.49	0.51	0.53	0.41	0.43	0.54	0.51	0.57	0.53
QWK	0.69	0.77	0.73	0.60	0.62	0.70	0.71	0.73	0.71
Pearson *r*	0.70	0.78	0.74	0.60	0.63	0.70	0.72	0.74	0.71

Note. M1=marker 1; M2=marker 2; C=CRASE

Table 14. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for N3_357

	Character and Setting			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	154	154	154	154	154	154	154	154	154
Score Distributions									
0	0.7%	1%	1%	0.7%	1%	0.7%	0.7%	1%	0 %
1	25%	22%	21%	8%	5%	6%	11%	10%	10%
2	53%	57%	56%	62%	62%	68%	77%	72%	76%
3	20%	18%	21%	29%	27%	21%	10%	15%	14%
4	2%	1%	0 %	0 %	3%	5%	0.7%	1%	0.7%
5				0.65%	0.7%	0 %			
6									
Mean	1.98	1.96	1.98	2.22	2.28	2.22	1.99	2.05	2.05
SD	0.75	0.71	0.69	0.64	0.70	0.66	0.52	0.60	0.51
$d_{M3\text{-}C}$			0.00			0.00			0.12
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	68%	71%	75%	73%	72%	73%	71%	76%	71%
Adjacent	31%	29%	25%	25%	28%	26%	29%	24%	28%
Non-Adjacent	2%	0%	0%	1%	0%	0.7%	0.7%	0%	0.7%
Kappa	0.47	0.51	0.59	0.50	0.46	0.49	0.30	0.43	0.26
QWK	0.64	0.70	0.76	0.66	0.70	0.66	0.50	0.61	0.42
Pearson *r*	0.64	0.70	0.76	0.66	0.70	0.66	0.51	0.62	0.42

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 15. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for N3_357

	Paragraphing			Sentence Structure			Punctuation		
	M1	**M2**	**C**	**M1**	**M2**	**C**	**M1**	**M2**	**C**
N	154	154	154	154	154	154	154	154	154
Score Distributions									
0	37%	37%	38%	1%	1%	**0%**	2%	3%	3%
1	58%	56%	56%	6%	6%	8%	37%	36%	35%
2	5%	7%	6%	49%	43%	54%	40%	37%	41%
3				27%	31%	25%	20%	21%	17%
4				16%	16%	13%	1%	3%	5%
5				0.7%	3%	0.7%	**0%**	**0%**	**0%**
6				**0%**	**0%**	**0%**			
Mean	0.68	0.70	0.69	2.52	2.62	2.45	1.82	1.83	1.86
SD	0.57	0.60	0.59	0.90	0.97	0.84	0.82	0.88	0.89
$d_{M3\text{-}C}$			0.02			-0.08			0.05
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	88%	73%	**76%**	52%	60%	63%	72%	67%	**66%**
Adjacent	12%	27%	24%	45%	36%	34%	27%	32%	34%
Non-Adjacent	0%	0%	0%	3%	4%	3%	0.7%	0.7%	0.7%
Kappa	0.77	0.50	0.55	0.29	0.40	0.42	0.59	0.52	0.49
QWK	0.82	0.61	**0.64**	0.67	0.69	0.70	0.79	0.78	0.75
Pearson *r*	0.82	0.61	0.64	0.67	0.70	0.71	0.80	0.78	0.75

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 16. Marker and CRASE agreement for spelling for N3_357

	Spelling		
	M1	**M2**	**C**
N	154	154	154
Score Distributions			
0	0.7%	0.7%	0.7%
1	6%	5%	3%
2	19%	16%	20%
3	40%	38%	40%
4	26%	30%	25%
5	7%	9%	10%
6	0.7%	0.7%	**0 %**
Mean	3.08	3.21	3.18
SD	1.06	1.05	1.02
d_{M3-C}			0.10
Agreement Indices			
	M1–M2	**C-M2**	**C–M1**
Exact	56%	57%	58%
Adjacent	42%	42%	41%
Non-Adjacent	1%	1%	0.7%
Kappa	0.40	0.41	0.43
QWK	0.79	0.78	0.80
Pearson *r*	0.79	0.78	0.80

Note. M1=marker 1; M2=marker 2; C=CRASE.

Prompt P6_579

Table 17. Marker and CRASE agreement for audience, text structure and ideas for P6_579

	Audience			Text Structure			Ideas		
	M1	M2	C	M1	M2	C	M1	M2	C
N	172	172	172	172	172	172	172	172	172
Score Distributions									
0	0.6%	0.6%	1%	1%	1%	1%	0.6%	0.6%	1%
1	2%	2%	0.6%	17%	20%	22%	2%	1%	1%
2	20%	18%	19%	47%	44%	42%	25%	21%	22%
3	41%	45%	40%	28%	31%	32%	53%	57%	55%
4	26%	25%	30%	6%	4%	3%	17%	19%	22%
5	9%	9%	9%				3%	1%	**0%**
6	2%	1%	**0%**						
Mean	3.24	3.23	3.25	2.21	2.17	2.15	2.92	2.97	2.94
SD	1.04	0.98	0.97	0.85	0.83	0.84	0.81	0.74	0.76
$d_{M3\text{-}C}$			0.01			-0.07			0.03
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	64%	60%	60%	62%	59%	63%	66%	70%	61%
Adjacent	34%	38%	38%	35%	39%	35%	33%	30%	39%
Non-Adjacent	2%	2%	2%	2%	2%	2%	1%	0%	0%
Kappa	0.49	0.43	0.44	0.43	0.39	0.46	0.44	0.50	0.37
QWK	0.80	0.76	0.77	0.68	0.67	0.70	0.68	0.73	0.68
Pearson *r*	0.80	0.76	0.77	0.68	0.67	0.71	0.69	0.73	0.68

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 18. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P6_579

	Persuasive Devices			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	172	172	172	172	172	172	172	172	172
Score Distributions									
0	1%	2%	2%	0.6%	0.6%	1%	0.6%	0.6%	1%
1	7%	5%	6%	2%	2%	1%	4%	3%	4%
2	51%	48%	52%	42%	49%	43%	65%	58%	60%
3	34%	41%	36%	40%	40%	37%	24%	35%	31%
4	6%	5%	5%	13%	8%	18%	6%	3%	3%
5				2%	1%	0%			
6									
Mean	2.38	2.42	2.36	2.69	2.56	2.69	2.31	2.37	2.32
SD	0.76	0.73	0.74	0.83	0.74	0.82	0.68	0.64	0.66
$d_{M3\text{-}C}$			-0.03			0.00			0.01
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	71%	74%	70%	67%	66%	68%	67%	73%	75%
Adjacent	28%	26%	30%	31%	34%	31%	33%	27%	25%
Non-Adjacent	0.6%	0.6%	0%	1%	0.6%	0.6%	0.6%	0%	0%
Kappa	0.52	0.56	0.51	0.48	0.46	0.51	0.38	0.50	0.53
QWK	0.72	0.74	0.73	0.71	0.71	0.75	0.60	0.68	0.72
Pearson *r*	0.72	0.74	0.74	0.73	0.72	0.75	0.60	0.68	0.72

Note. M1=marker 1; M2=marker 2; C=CRASE

Table 19. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for P6_579

	Paragraphing			Sentence Structure			Punctuation		
	M1	M2	C	M1	M2	C	M1	M2	C
N	172	172	172	172	172	172	172	172	172
Score Distributions									
0	9%	7%	8%	1%	1%	1%	0.6%	0.6%	2%
1	48%	47%	48%	0.6%	0.6%	2%	12%	14%	12%
2	38%	42%	41%	23%	22%	20%	40%	33%	38%
3	5%	4%	3%	36%	44%	42%	39%	44%	38%
4				37%	27%	32%	8%	8%	8%
5				1%	5%	3%	0.6%	1%	2%
6				0.6%	0.6%	0%			
Mean	1.39	1.43	1.40	3.13	3.12	3.12	2.44	2.48	2.42
SD	0.73	0.68	0.69	0.91	0.92	0.91	0.85	0.90	0.95
$d_{M3\text{-}C}$			0.01			-0.01			-0.02
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	68%	64%	70%	60%	58%	62%	66%	53%	56%
Adjacent	32%	36%	30%	38%	41%	37%	34%	44%	41%
Non-Adjacent	0%	0%	0%	2%	1%	0.6%	0.6%	2%	3%
Kappa	0.47	0.40	0.51	0.42	0.39	0.45	0.49	0.32	0.35
QWK	0.68	0.62	0.70	0.73	0.73	0.76	0.76	0.68	0.66
Pearson *r*	0.68	0.62	0.70	0.73	0.73	0.76	0.77	0.69	0.67

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 20. Marker and CRASE agreement for spelling for P6_579

	Spelling		
	M1	**M2**	**C**
N	172	172	172
Score Distributions			
0	0.6%	0.6%	1%
1	1%	2%	1%
2	5%	6%	1%
3	27%	31%	33%
4	39%	35%	37%
5	26%	25%	25%
6	1%	0.6%	2%
Mean	3.86	3.76	3.86
SD	0.98	1.01	0.98
$\mathbf{d_{M3\text{-}C}}$			0.00
Agreement Indices			
	M1–M2	**C-M2**	**C–M1**
Exact	62%	63%	69%
Adjacent	38%	35%	30%
Non-Adjacent	0.6%	2%	1%
Kappa	0.46	0.48	0.56
QWK	0.80	0.79	0.82
Pearson *r*	0.80	0.79	0.82

Note. M1=marker 1; M2=marker 2; C=CRASE.

Prompt P7_579

Table 21. Marker and CRASE agreement for audience, text structure and ideas for P7_579

	Audience			Text Structure			Ideas		
	M1	M2	C	M1	M2	C	M1	M2	C
N	158	158	158	158	158	158	158	158	158
Score Distributions									
0	0%	0%	0%	0%	0.6%	0.6%	0%	0%	0.6%
1	0.6%	2%	2%	19%	16%	20%	1%	3%	1%
2	13%	11%	17%	47%	42%	44%	16%	11%	19%
3	47%	46%	40%	28%	34%	30%	58%	60%	53%
4	30%	27%	31%	6%	6%	6%	23%	25%	25%
5	8%	13%	10%				1%	1%	0.6%
6	1%	1%	0%						
Mean	3.36	3.41	3.30	2.21	2.29	2.22	3.08	3.11	3.03
SD	0.88	0.96	0.94	0.81	0.84	0.86	0.70	0.71	0.76
$d_{M3\text{-}C}$			-0.07			0.01			-0.07
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	60%	65%	67%	58%	63%	62%	70%	70%	68%
Adjacent	38%	34%	31%	40%	35%	37%	30%	29%	31%
Non-Adjacent	2%	1%	2%	2%	1%	0.6%	0%	0.6%	0.6%
Kappa	0.41	0.50	0.52	0.38	0.46	0.43	0.48	0.50	0.47
QWK	0.73	0.78	0.76	0.65	0.72	0.71	0.70	0.71	0.68
Pearson *r*	0.73	0.79	0.77	0.66	0.72	0.71	0.70	0.71	0.69

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 22. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P7_579

	Persuasive Devices			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	158	158	158	158	158	158	158	158	158
Score Distributions									
0	**0%**	0.6%	0.6%	**0%**	**0%**	0.6%	**0%**	**0%**	0.6%
1	9%	8%	6%	0.6%	0.6%	2%	4%	2%	1%
2	48%	42%	52%	44%	46%	40%	59%	59%	58%
3	36%	43%	35%	37%	35%	39%	33%	35%	37%
4	6%	6%	6%	17%	16%	18%	4%	4%	4%
5				1%	2%	**0%**			
6									
Mean	2.39	2.45	2.40	2.75	2.73	2.73	2.38	2.41	2.42
SD	0.75	0.75	0.73	0.79	0.81	0.80	0.64	0.60	0.62
d_{M3-C}			0.01			-0.03			0.06
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	66%	68%	66%	57%	59%	61%	70%	77%	72%
Adjacent	34%	32%	32%	39%	41%	37%	30%	22%	25%
Non-Adjacent	0.6%	0.6%	2%	4%	0%	1%	0.6%	0.6%	3%
Kappa	0.46	0.48	0.44	0.33	0.37	0.40	0.43	0.57	0.48
QWK	0.68	0.69	0.63	0.57	0.68	0.66	0.57	0.67	0.55
Pearson *r*	0.68	0.69	0.63	0.57	0.68	0.66	0.58	0.67	0.55

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 23. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for P7_579

	Paragraphing			Sentence Structure			Punctuation		
	M1	M2	C	M1	M2	C	M1	M2	C
N	158	158	158	158	158	158	158	158	158
Score Distributions									
0	8%	8%	11%	**0.0%**	**0.0%**	0.6%	0.6%	1%	1%
1	49%	45%	51%	2%	0.6%	1%	11%	11%	9%
2	37%	39%	31%	26%	25%	21%	34%	28%	35%
3	6%	8%	7%	34%	36%	44%	42%	43%	44%
4				33%	35%	25%	11%	17%	10%
5				5%	3%	8%	1%	**0%**	**0%**
6				0.6%	**0%**	**0%**			
Mean	1.42	1.48	1.34	3.15	3.15	3.15	2.55	2.64	2.53
SD	0.73	0.75	0.77	0.95	0.86	0.93	0.90	0.93	0.85
d_{M3-C}			-0.11			0.00			-0.02
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	66%	66%	**59%**	61%	63%	**51%**	61%	56%	56%
Adjacent	34%	34%	41%	37%	34%	46%	39%	43%	42%
Non-Adjacent	0.6%	0%	0.6%	1%	3%	3%	0%	1%	2%
Kappa	0.45	0.47	0.34	0.45	0.47	0.31	0.43	0.35	0.35
QWK	0.67	0.72	0.62	0.74	0.71	**0.63**	0.77	0.70	0.68
Pearson *r*	0.67	0.73	0.62	0.74	0.71	0.63	0.77	0.71	0.68

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 24. Marker and CRASE agreement for spelling for P7_579

	Spelling		
	M1	**M2**	**C**
N	158	158	158
Score Distributions			
0	**0 %**	**0 %**	0.6%
1	0.6%	1%	0.6%
2	4%	3%	4%
3	20%	25%	18%
4	38%	35%	35%
5	35%	33%	39%
6	3%	3%	3%
Mean	4.11	4.06	4.12
SD	0.92	0.96	1.00
$d_{M3\text{-}C}$			0.01
Agreement Indices			
	M1–M2	**C-M2**	**C–M1**
Exact	69%	66%	65%
Adjacent	31%	33%	34%
Non-Adjacent	0%	0.6%	2%
Kappa	0.56	0.52	0.49
QWK	0.82	0.81	0.78
Pearson *r*	0.83	0.82	0.78

Note. M1=marker 1; M2=marker 2; C=CRASE.

Prompt P8_579

Table 25. Marker and CRASE agreement for audience, text structure and ideas for P8_579

	Audience			Text Structure			Ideas		
	M1	**M2**	**C**	**M1**	**M2**	**C**	**M1**	**M2**	**C**
N	138	138	138	138	138	138	138	138	138
Score Distributions									
0	**0%**	**0%**	**0%**	0.7%	**0%**	**0%**	**0%**	**0%**	**0%**
1	1%	2%	0.7%	26%	24%	30%	1%	3%	0.7%
2	17%	17%	14%	40%	47%	37%	28%	30%	22%
3	43%	47%	44%	27%	24%	30%	48%	47%	57%
4	28%	25%	30%	7%	5%	4%	17%	16%	20%
5	9%	7%	11%				6%	4%	**0%**
6	1%	1%	**0%**						
Mean	3.29	3.21	3.36	2.12	2.10	2.07	2.97	2.87	2.97
SD	0.95	0.93	0.89	0.90	0.82	0.86	0.86	0.84	0.67
$d_{M3\text{-}C}$			0.08			-0.06			0.00
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	67%	63%	67%	66%	64%	68%	63%	65%	64%
Adjacent	31%	36%	33%	34%	34%	30%	36%	35%	36%
Non-Adjacent	1%	0.7%	0.7%	0%	1%	1%	0.7%	0%	0%
Kappa	0.53	0.46	0.52	0.50	0.48	0.54	0.44	0.45	0.43
QWK	0.79	0.76	0.79	0.77	0.69	0.74	0.73	0.70	0.69
Pearson *r*	0.79	0.78	0.79	0.77	0.69	0.74	0.74	0.73	0.72

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 26. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for P8_579

	Persuasive Devices			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	138	138	138	138	138	138	138	138	138
Score Distributions									
0	0.7%	0 %	0 %	0 %	0 %	0 %	0.7%	0 %	0 %
1	7%	7%	6%	1%	1%	0.7%	4%	6%	2%
2	47%	58%	52%	45%	46%	53%	59%	63%	59%
3	36%	31%	38%	36%	34%	30%	29%	27%	38%
4	9%	4%	4%	13%	17%	15%	7%	4%	0 %
5				4%	2%	0.7%			
6									
Mean	2.45	2.33	2.40	2.74	2.73	2.62	2.38	2.30	2.36
SD	0.78	0.67	0.66	0.87	0.84	0.78	0.71	0.64	0.53
d_{M3-C}			-0.07			-0.15			-0.03
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	66%	75%	70%	60%	64%	69%	70%	72%	75%
Adjacent	33%	25%	30%	39%	35%	29%	30%	28%	25%
Non-Adjacent	0.7%	0%	0%	0.7%	0.7%	2%	0%	0%	0%
Kappa	0.44	0.56	0.51	0.39	0.44	0.51	0.44	0.46	0.53
QWK	0.66	0.71	0.71	0.71	0.71	0.72	0.67	0.59	0.67
Pearson *r*	0.68	0.71	0.73	0.71	0.72	0.73	0.68	0.61	0.70

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 27. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for P8_579

	Paragraphing			Sentence Structure			Punctuation		
	M1	M2	C	M1	M2	C	M1	M2	C
N	138	138	138	138	138	138	138	138	138
Score Distributions									
0	16%	11%	10%	0 %	0 %	0 %	0.7%	0.7%	0 %
1	41%	48%	50%	1%	4%	1%	14%	14%	14%
2	35%	38%	38%	32%	22%	29%	38%	29%	33%
3	8%	4%	2%	32%	42%	33%	34%	46%	40%
4				28%	29%	33%	13%	11%	13%
5				7%	2%	3%	0 %	0 %	0 %
6				0.7%	1%	0.7%			
Mean	1.35	1.34	1.32	3.08	3.09	3.09	2.45	2.52	2.51
SD	0.84	0.72	0.68	0.89	0.93	0.93	0.91	0.89	0.90
d_{M3-C}			-0.04			0.01			0.07
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	63%	70%	64%	58%	59%	65%	59%	62%	60%
Adjacent	37%	30%	35%	40%	37%	32%	39%	37%	38%
Non-Adjacent	0%	0.7%	0.7%	2%	4%	3%	2%	1%	1%
Kappa	0.43	0.50	0.45	0.41	0.41	0.51	0.41	0.44	0.43
QWK	0.70	0.67	0.68	0.74	0.68	0.76	0.70	0.73	0.73
Pearson *r*	0.71	0.67	0.69	0.74	0.68	0.76	0.71	0.73	0.73

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 28. Marker and CRASE Agreement for Spelling for P8_579

	Spelling		
	M1	**M2**	**C**
N	138	138	138
Score Distributions			
0	**0 %**	**0 %**	**0 %**
1	0.7%	0.7%	0.7%
2	12%	8%	6%
3	20%	28%	27%
4	37%	29%	33%
5	28%	35%	33%
6	3%	**0 %**	0.7%
Mean	3.87	3.89	3.94
SD	1.07	1.00	0.96
$\mathbf{d_{M3\text{-}C}}$			0.07
Agreement Indices			
	M1–M2	**C-M2**	**C–M1**
Exact	64%	70%	66%
Adjacent	36%	29%	33%
Non-Adjacent	0%	0.7%	1%
Kappa	0.52	0.58	0.53
QWK	0.83	0.83	0.81
Pearson *r*	0.83	0.83	0.82

Note. M1=marker 1; M2=marker 2; C=CRASE.

Prompt N4_579

Table 29. Marker and CRASE agreement for audience, text structure and ideas for N4_579

	Audience			Text Structure			Ideas		
	M1	**M2**	**C**	**M1**	**M2**	**C**	**M1**	**M2**	**C**
N	138	138	138	138	138	138	138	138	138
Score Distributions									
0	**0%**	**0%**	**0%**	0.7%	**0%**	**0%**	**0%**	**0%**	**0%**
1	1%	**0%**	0.7%	13%	17%	13%	3%	1%	3%
2	20%	17%	20%	54%	53%	50%	28%	25%	27%
3	43%	48%	46%	22%	26%	35%	51%	57%	56%
4	24%	25%	26%	9%	4%	2%	17%	17%	14%
5	8%	9%	8%				2%	0.7%	**0%**
6	4%	0.7%	**0%**						
Mean	3.30	3.28	3.21	2.27	2.18	2.26	2.88	2.91	2.82
SD	1.06	0.88	0.88	0.83	0.76	0.71	0.80	0.70	0.71
d_{M3-C}			-0.09			-0.01			-0.08
Agreement Indices									
	M1–M2	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**	**M1–M2**	**C-M2**	**C–M1**
Exact	59%	57%	**51%**	58%	57%	**51%**	65%	69%	72%
Adjacent	39%	43%	43%	35%	41%	44%	33%	30%	28%
Non-Adjacent	1%	0.7%	6%	7%	2%	4%	1%	0.7%	0.7%
Kappa	0.41	0.36	0.29	0.33	0.30	0.23	0.44	0.48	0.54
QWK	0.76	0.70	**0.63**	0.50	0.53	0.48	0.65	0.66	0.73
Pearson *r*	0.77	0.70	0.64	0.50	0.54	0.49	0.66	0.67	0.74

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 30. Marker and CRASE agreement for persuasive devices, vocabulary and cohesion for N4_579

	Character and Setting			Vocabulary			Cohesion		
	M1	M2	C	M1	M2	C	M1	M2	C
N	138	138	138	138	138	138	138	138	138
Score Distributions									
0	0%	0%	0%	0%	0%	0%	0%	0%	0%
1	3%	3%	4%	0%	0.7%	0.7%	2%	0.7%	3%
2	47%	51%	46%	52%	45%	44%	73%	75%	70%
3	43%	40%	47%	33%	41%	41%	19%	22%	27%
4	7%	7%	3%	12%	12%	14%	6%	2%	0.7%
5				4%	2%	0.7%			
6									
Mean	2.54	2.50	2.49	2.67	2.70	2.70	2.28	2.26	2.25
SD	0.67	0.67	0.62	0.82	0.77	0.74	0.60	0.50	0.51
$d_{M3\text{-}C}$			-0.08			0.04			-0.05
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	70%	73%	70%	58%	66%	62%	71%	76%	67%
Adjacent	30%	27%	30%	42%	33%	36%	28%	24%	30%
Non-Adjacent	0%	0%	0%	0%	0.7%	2%	0.7%	0%	2%
Kappa	0.48	0.53	0.49	0.32	0.45	0.38	0.29	0.43	0.26
QWK	0.66	0.67	0.64	0.67	0.68	0.63	0.49	0.53	**0.37**
Pearson *r*	0.66	0.67	0.65	0.67	0.68	0.63	0.50	0.53	0.38

Note. M1=marker 1; M2=marker **2**; C=CRASE.

Table 31. Marker and CRASE agreement for paragraphing, sentence structure and punctuation for N4_579

	Paragraphing			Sentence Structure			Punctuation		
	M1	M2	C	M1	M2	C	M1	M2	C
N	138	138	138	138	138	138	138	138	138
Score Distributions									
0	12%	15%	17%	0%	0%	0%	3%	4%	3%
1	78%	75%	78%	0%	0%	3%	14%	14%	13%
2	10%	9%	5%	28%	21%	25%	38%	31%	38%
3				38%	43%	41%	39%	43%	41%
4				26%	30%	30%	4%	8%	4%
5				7%	6%	1%	1%	0%	1%
6				1%	0%	0%			
Mean	0.98	0.94	0.88	3.15	3.21	3.01	2.32	2.37	2.35
SD	0.48	0.50	0.46	0.96	0.84	0.85	0.93	0.95	0.92
$d_{M3\text{-}C}$			**-0.21**			-0.15			0.03
Agreement Indices									
	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1	M1–M2	C-M2	C–M1
Exact	80%	79%	**74%**	57%	55%	**46%**	59%	58%	56%
Adjacent	20%	21%	26%	39%	40%	47%	38%	40%	42%
Non-Adjacent	0%	0%	0%	4%	5%	7%	2%	2%	2%
Kappa	0.49	0.45	0.30	0.39	0.34	0.22	0.41	0.39	0.35
QWK	0.58	0.54	**0.41**	0.67	0.59	**0.54**	0.73	0.72	0.70
Pearson *r*	0.58	0.55	0.42	0.68	0.61	0.55	0.73	0.72	0.70

Note. M1=marker 1; M2=marker 2; C=CRASE.

Table 32. Marker and CRASE agreement for spelling for N4_579

	Spelling		
	M1	**M2**	**C**
N	138	138	138
Score Distributions			
0	**0 %**	**0 %**	**0 %**
1	**0 %**	0.7%	0.7%
2	9%	6%	9%
3	22%	27%	21%
4	40%	39%	38%
5	26%	26%	30%
6	3%	1%	0.7%
Mean	3.91	3.88	3.90
SD	0.99	0.94	0.99
$\mathbf{d_{M3\text{-}C}}$			-0.01
Agreement Indices			
	M1–M2	**C-M2**	**C–M1**
Exact	66%	61%	64%
Adjacent	33%	39%	35%
Non-Adjacent	0.7%	0%	0.7%
Kappa	0.52	0.45	0.50
QWK	0.80	0.79	0.81
Pearson *r*	0.80	0.79	0.81

Note. M1=marker 1; M2=marker 2; C=CRASE.

Intra-criteria correlations

Prompt P3_357

Table 33. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for P3_357

Source	N	Mean	SD	Min	Max
Marker 1	142	0.686	0.074	0.529	0.888
CRASE	142	0.783	0.059	0.643	0.9

Table 34. Intra-criteria correlations of CRASE (top), M1 (bottom) for P3_357

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.82	0.86	0.85	0.81	0.78	0.82	0.90	0.76	0.84
TS	0.69		0.81	0.76	0.80	0.67	0.80	0.85	0.74	0.79
ID	0.89	0.65		0.73	0.82	0.68	0.77	0.85	0.72	0.82
PD	0.77	0.62	0.74		0.73	0.77	0.78	0.84	0.70	0.81
VO	0.77	0.61	0.73	0.70		0.69	0.78	0.81	0.73	0.85
CO	0.75	0.57	0.74	0.74	0.70		0.64	0.77	0.69	0.74
PA	0.73	0.69	0.71	0.71	0.61	0.64		0.84	0.73	0.84
SS	0.81	0.60	0.80	0.75	0.72	0.73	0.70		0.79	0.85
PU	0.63	0.53	0.60	0.61	0.58	0.59	0.63	0.69		0.78
SP	0.74	0.57	0.71	0.73	0.71	0.64	0.68	0.72	0.65	

Prompt P4_357

Table 35. Summary statistics of the intra-criteria correlations ofM1, M2, M3 and CRASE for P4_357

Source	N	Mean	SD	Min	Max
Marker 1	125	0.728	0.066	0.549	0.884
CRASE	125	0.791	0.067	0.588	0.889

Table 36. Intra-criteria correlations of CRASE (top), M1 (bottom) for P4_357

	AU	TS	ID	CS	VO	CO	PA	SS	PU	SP
AU		0.89	0.88	0.82	0.80	0.82	0.84	0.89	0.74	0.87
TS	0.82		0.82	0.82	0.71	0.74	0.85	0.87	0.76	0.85
ID	0.88	0.81		0.79	0.68	0.76	0.80	0.82	0.74	0.84
CS	0.81	0.80	0.73		0.71	0.83	0.82	0.83	0.69	0.86
VO	0.80	0.74	0.73	0.73		0.78	0.73	0.76	0.59	0.79
CO	0.80	0.70	0.72	0.73	0.78		0.71	0.82	0.67	0.80
PA	0.81	0.77	0.74	0.72	0.76	0.76		0.85	0.77	0.86
SS	0.75	0.68	0.69	0.72	0.72	0.76	0.72		0.78	0.87
PU	0.66	0.65	0.61	0.65	0.55	0.60	0.64	0.65		0.77
SP	0.79	0.71	0.79	0.70	0.71	0.77	0.73	0.68	0.65	

Prompt P5_357

Table 37. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for P5_357

Source	N	Mean	SD	Min	Max
Marker 1	128	0.741	0.054	0.638	0.877
CRASE	128	0.785	0.063	0.666	0.934

Table 39. Intra-criteria correlations of CRASE (top), M1 (bottom) for P5_357

	AU	TS	ID	CS	VO	CO	PA	SS	PU	SP
AU		0.90	0.84	0.80	0.73	0.79	0.87	0.93	0.81	0.87
TS	0.84		0.76	0.78	0.74	0.69	0.87	0.87	0.80	0.83
ID	0.88	0.78		0.71	0.67	0.72	0.75	0.83	0.79	0.81
CS	0.78	0.76	0.77		0.73	0.78	0.80	0.78	0.76	0.81
VO	0.84	0.69	0.79	0.72		0.70	0.72	0.72	0.68	0.71
CO	0.71	0.72	0.74	0.74	0.70		0.71	0.81	0.77	0.74
PA	0.78	0.82	0.73	0.71	0.71	0.69		0.84	0.78	0.81
SS	0.85	0.80	0.79	0.75	0.79	0.72	0.75		0.84	0.88
PU	0.70	0.70	0.71	0.70	0.68	0.64	0.68	0.74		0.82
SP	0.79	0.69	0.75	0.74	0.74	0.68	0.67	0.75	0.66	

Prompt N3_357

Table 40. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for N3_357

Source	N	Mean	SD	Min	Max
Marker 1	154	0.621	0.099	0.428	0.775
CRASE	154	0.763	0.078	0.566	0.902

Table 41. Intra-criteria correlations of CRASE (top), M1 (bottom) for N3_357

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.81	0.83	0.77	0.81	0.76	0.69	0.90	0.76	0.84
TS	0.77		0.73	0.84	0.73	0.71	0.79	0.79	0.76	0.84
ID	0.78	0.69		0.76	0.86	0.71	0.57	0.86	0.73	0.83
PD	0.74	0.71	0.73		0.78	0.69	0.69	0.79	0.73	0.84
VO	0.69	0.59	0.69	0.65		0.78	0.57	0.86	0.73	0.79
CO	0.63	0.61	0.60	0.58	0.64		0.60	0.82	0.71	0.77
PA	0.57	0.53	0.53	0.49	0.50	0.44		0.64	0.66	0.72
SS	0.77	0.64	0.73	0.70	0.72	0.71	0.48		0.82	0.86
PU	0.59	0.54	0.56	0.55	0.45	0.43	0.45	0.64		0.79
SP	0.75	0.66	0.70	0.69	0.70	0.55	0.54	0.68	0.57	

Prompt P6_579

Table 42. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for P6_579

Source	N	Mean	SD	Min	Max
Marker 1	172	0.724	0.077	0.549	0.868
CRASE	172	0.825	0.045	0.735	0.905

Table 43. Intra-criteria correlations of CRASE (top), M1 (bottom) for P6_579

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.90	0.88	0.87	0.86	0.85	0.84	0.88	0.80	0.87
TS	0.84		0.86	0.85	0.81	0.82	0.85	0.88	0.81	0.84
ID	0.86	0.80		0.76	0.82	0.78	0.75	0.86	0.77	0.86
PD	0.80	0.79	0.77		0.81	0.89	0.85	0.86	0.74	0.83
VO	0.87	0.76	0.81	0.79		0.81	0.78	0.83	0.77	0.87
CO	0.78	0.75	0.74	0.75	0.74		0.78	0.87	0.75	0.83
PA	0.77	0.75	0.72	0.73	0.73	0.71		0.81	0.73	0.77
SS	0.74	0.72	0.68	0.74	0.73	0.68	0.67		0.84	0.87
PU	0.65	0.62	0.56	0.58	0.62	0.55	0.55	0.66		0.77
SP	0.79	0.73	0.73	0.78	0.78	0.67	0.67	0.76	0.66	

Prompt P7_579

Table 44. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for P7_579

Source	N	Mean	SD	Min	Max
Marker 1	158	0.686	0.08	0.502	0.858
CRASE	158	0.817	0.044	0.717	0.898

Table 45. Intra-criteria correlations of CRASE (top), M1 (Bottomb) for P7_579

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.90	0.86	0.87	0.85	0.87	0.87	0.88	0.79	0.84
TS	0.81		0.85	0.84	0.82	0.83	0.86	0.89	0.80	0.80
ID	0.86	0.73		0.79	0.82	0.77	0.81	0.82	0.74	0.77
PD	0.85	0.75	0.77		0.79	0.88	0.84	0.83	0.77	0.77
VO	0.81	0.70	0.75	0.76		0.80	0.79	0.83	0.76	0.80
CO	0.78	0.65	0.69	0.74	0.71		0.80	0.80	0.72	0.72
PA	0.72	0.71	0.66	0.69	0.66	0.62		0.87	0.77	0.79
SS	0.74	0.66	0.71	0.71	0.64	0.73	0.66		0.85	0.85
PU	0.61	0.57	0.64	0.57	0.50	0.56	0.51	0.62		0.81
SP	0.70	0.64	0.72	0.67	0.65	0.61	0.61	0.69	0.70	

Prompt P8_579

Table 46. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for P8_579

Source	N	Mean	SD	Min	Max
Marker 1	138	0.76	0.064	0.623	0.872
CRASE	138	0.789	0.053	0.683	0.892

Table 47. Intra-criteria correlations of CRASE (top), M1 (bottom) for P8_579

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.82	0.84	0.87	0.83	0.85	0.81	0.86	0.81	0.82
TS	0.87		0.77	0.78	0.71	0.80	0.82	0.82	0.76	0.85
ID	0.87	0.81		0.72	0.71	0.69	0.69	0.78	0.76	0.75
PD	0.85	0.81	0.78		0.81	0.87	0.80	0.85	0.78	0.82
VO	0.84	0.81	0.84	0.78		0.82	0.68	0.82	0.73	0.76
CO	0.84	0.76	0.78	0.84	0.81		0.79	0.81	0.75	0.78
PA	0.82	0.79	0.72	0.75	0.74	0.70		0.72	0.71	0.77
SS	0.81	0.79	0.79	0.73	0.80	0.75	0.78		0.82	0.89
PU	0.70	0.69	0.68	0.63	0.67	0.62	0.69	0.77		0.80
SP	0.78	0.71	0.73	0.73	0.71	0.65	0.72	0.74	0.71	

Prompt N4_579

Table 48. Summary statistics of the intra-criteria correlations of M1, M2, M3, and CRASE for N4_579

Source	N	Mean	SD	Min	Max
Marker 1	138	0.592	0.116	0.386	0.842
CRASE	138	0.717	0.094	0.505	0.865

Table 49. Intra-criteria correlations of CRASE (top), M1 (bottom) for N4_579

	AU	TS	ID	PD	VO	CO	PA	SS	PU	SP
AU		0.84	0.82	0.78	0.81	0.71	0.66	0.86	0.66	0.81
TS	0.72		0.74	0.79	0.74	0.62	0.64	0.81	0.57	0.78
ID	0.84	0.69		0.74	0.77	0.67	0.65	0.84	0.64	0.82
PD	0.77	0.64	0.74		0.82	0.68	0.60	0.80	0.57	0.78
VO	0.81	0.58	0.73	0.67		0.72	0.51	0.80	0.62	0.78
CO	0.62	0.62	0.60	0.53	0.62		0.50	0.77	0.66	0.71
PA	0.51	0.44	0.48	0.42	0.39	0.45		0.64	0.60	0.61
SS	0.73	0.63	0.68	0.62	0.68	0.69	0.44		0.73	0.85
PU	0.57	0.43	0.55	0.51	0.50	0.44	0.43	0.65		0.72
SP	0.68	0.50	0.67	0.60	0.62	0.54	0.39	0.62	0.61	

Summed Scores Correlations

Table 26. QWK of summed scored and difference in QWK

Writing Prompt	Marker 1 Mean (SD)	CRASE Mean (SD)	Marker 1—Marker2 QWK	CRASE-Marker 1 QWK
P3_357	21.80 (6.66)	21.79 (7.01)	0.92	0.88
P4_357	22.10 (7.43)	22.46 (7.48)	0.93	0.93
P5_357	21.59 (7.17)	21.08 (6.86)	0.92	0.91
N3_357	21.12 (6.10)	21.14 (6.60)	0.87	0.87
P6_579	26.58 (7.33)	26.51 (7.65)	0.89	0.87
P7_579	27.41 (6.82)	27.23 (7.55)	0.87	0.83
P8_579	26.70 (7.86)	26.64 (7.08)	0.88	0.88
N4_579	26.30 (6.57)	25.87 (6.44)	0.87	0.81

Appendix 2

The following modifications were made to the sample of narrative and persuasive scripts to generate a set of modified scripts used in the resilience analyses presented in study 2.

Lexical field

1. Insert replacement nouns and verbs unrelated to the prompt topic’s lexical theme.
2. Repeat topic word and words related to topic (use more topic-related words).
3. Adjust text so only one topic word is used.
4. Substitute a frequently used topic word with another word throughout the essay.
5. Insert a paragraph with content about an unrelated topic.
6. Insert two random paragraphs with content about an unrelated topic.
7. Double the length of the essay with syntactically correct gobbledygook.
8. In a short text, repeat the text multiple times.

Sentence construction

1. Change the tense from past to present and present to past.
2. Mix tense throughout script.
3. Change mood: change high-intensity modals to low-intensity modals.
4. Add subordinate clauses to extend sentences.
5. Remove subordinate clauses to shorten sentences.

Phrase and clause connectors

1. Insert language associated with genre (e.g., in persuasive, insert extra connectives (words and phrases) to signal causal and conditional logic; in narrative, insert vivid description through use of additional adjective and adverbial words, phrases and clauses).
2. Insert words that signal the start of a point of an argument (persuasive).
3. Remove words that signal the start of a point of an argument (persuasive).

Punctuation

1. Improve sentence punctuation by correcting/adding sentence punctuation.
2. Remove sentence punctuation.
3. Remove all internal punctuation.
4. Adjust text so it contains only correct sentence punctuation and no internal punctuation.
5. Add internal punctuation so that it interferes with smooth reading.
6. Insert a large amount of high-level punctuation – use every marker known to people.

Length – adapt by repeating or removing sections of text

1. High-quality long essay – double the length.
2. Low-quality long essay – double the length.
3. High-quality long essay – halve the length.

4. Low-quality long essay – half the length.
5. High-quality short essay – double the length.
6. Low-quality short essay – double the length.
7. High-quality short essay – halve the length.
8. Low-quality short essay – half the length.