AN EVALUATION OF AUTOMATED SCORING OF NAPLAN PERSUASIVE WRITING

Technical Report

2015

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

The National Assessment and Surveys Online Program (NASOP), funded by the Australian Government, was designed to support the delivery of the National Assessment Program (NAPLAN and NAP Sample) online. ACARA developed a comprehensive research program to address a range of transition issues, including test design and impacts on student performance, measurement and reporting.

To explore the possibilities that digital technologies offer in providing faster feedback on student performance in NAPLAN writing tests, ACARA conducted a pilot research study to explore the capacity of automated essay scoring systems to reliably mark NAPLAN persuasive essays using the NAPLAN persuasive writing marking rubric.

Four vendors were independently engaged to score the NAPLAN persuasive essays. Each utilised a different proprietary automated scoring system:

- Measurement Incorporated – Project Essays Grader (PEG)
- Pearson – Intelligent Essay Assessor (IEA)
- Pacific Metrics – Constructed-Response Automated Scoring Engine (CRASE®)
- MetaMetrics – Lexile® Writing Analyzer

The outcome of the research is summarised in the research report, *An Evaluation of Automated Scoring of NAPLAN Persuasive Writing.* This technical report is a compilation of the vendor reports that contain more detailed information about the performance of the four automated scoring systems in the marking of NAPLAN persuasive writing.

Australian Curriculum, Assessment, and Reporting Authority (ACARA)





NAPLAN Online Trial Study 2013: Automated Essay Scoring of Writing Scripts Final Report

Shayne D. Miel, David Vaughn, Justin Fister, Robert T. Millard, Julie St. John, and Gwillim Law

September 6, 2013

MEASUREMENT INCORPORATED
EXCELLENCE IN ASSESSMENT™
www.measurementinc.com

EXCELLENCE IN ASSESSMENT

TABLE OF CONTENTS

Executive Summary

The NAPLAN Online Trial Study 2013 – Automated Essay Scoring of Writing Scripts, was a joint effort between Measurement Incorporated (MI) and the Australian Curriculum, Assessment and Reporting Authority (ACARA) to analyze how well Project Essay Grade (PEG™), MI's automated scoring software, would perform on the NAPLAN writing task. ACARA provided MI with 1356 responses written to a prompt asking for persuasive essays. Of those responses, 1017 had been scored on 10 criteria by two independent readers, while 339 were left without scores. ACARA had divided the scored essays into two groups, a training set of 677 essays and a validation set of 340 essays. MI's task was to predict scores for the 10 criteria on the 339 unscored essays. PEG's performance was to be measured by how well the predicted scores agreed with the first of the two independent scores assigned by human readers. The second human score was to be used to evaluate how well two independent human readers could agree with one another, providing a baseline for the performance of the AI engine. Using PEG's advanced artificial intelligence software, models for each criterion were trained on the initial set of essays, results were validated, and scores were generated for the final, unscored set of 339 essays.

For all ten criteria that are part of the NAPLAN writing rubric, the PEG scores were equivalent to the scores assigned by human readers. Values for quadratic weighted kappa, Pearson's *r*, perfect agreement, and adjacent agreement demonstrated the efficacy of PEG as a means for providing reliable, accurate scores for student-produced essays. Final observations discuss the PEG results for the Cohesion and Punctuation criteria, with a discussion of the limits that human scoring places on the upper bounds of PEG models. Finally, a brief analysis using resolved scores – an average of human reader scores – is presented and suggestions are made for including this score in future studies.

MI's Artificial Intelligence Scoring Engine

Overview of Automated Scoring

MI has been at the forefront of scoring student writing since the early 1980s. MI pioneered many of the complex processes involved in handscoring student essays accurately and cost-efficiently—scoring numerous U.S. state departments of education writing assessments. By the late 1990s, MI's expertise in handscoring had firmly established the company as the industry's premier writing assessment company.

By early 2000, MI had also established a collegial relationship with Dr. Ellis Batten Page of nearby Duke University. Page, regarded as the "father of automated essay scoring" from his pioneering work in the early 1960s, was the first to explore, document, and validate the computer-based assessment of written prose. His software was entering a new era as advances in microcomputer technology and the emergence of the World Wide Web were making automated essay scoring a practical possibility. Eventually, in 2003, MI acquired the PEG technology from Dr. Page and his associates. Eleven years later, MI has re-engineered, enhanced and extended the PEG system using the latest techniques and technologies in the field of computational linguistics, machine learning, and natural language processing.

With subsequent improvements in PEG and general advances in the reliability of machine scoring, artificial intelligence (AI) scoring has become a valuable, and in some cases, essential, tool in a variety of contexts. MI's AI scoring engine, PEG, is currently in use in summative and formative assessments, and we are studying its use in computer adaptive assessments. It will be used in pilot and field tests for one of the major U. S. assessment consortia, where AI scoring will provide a necessary core element enabling the scoring of millions of student written responses. PEG, with an established track record in scoring essays for qualitative characteristics such as such as organization, support, word choice, and mechanics, has also performed well in studies of AI scoring for content and will be at the forefront of these national assessment developments.

Summative Assessments

Since 2009, the Utah State Office of Education has successfully used PEG as the sole scoring method on the statewide summative Direct Writing Assessment in Grades 5 and 8. Over the past four years, PEG has scored 344,000 student responses on Utah's six trait rubric. In addition, in 2013 PEG was used as the second reader on the Connecticut SBAC Aligned Practice Assessment (APA), providing scores for 90,000 student responses.

In Spring 2013, PEG was selected as one of the AI engines to be deployed by the Smarter Balanced Assessment Consortium (SBAC) to provide automated scoring of items on the pilot and field tests of its next generation assessments. Scoring models developed by MI's AI scoring engine will play a significant role in the automated scoring of extended essays and short constructed responses in SBAC-developed English Language Arts and Math assessments. Assessments on the scale proposed by SBAC could not exist without AI scoring.

In 2012, the Hewlett Foundation sponsored two global competitions in automated scoring – the Automated Student Assessment Prize (ASAP), Phases 1 and 2. These competitions were the first of their kind and were intended to independently evaluate the state of the art in essay and short answer scoring. MI is pleased to say that PEG took first place in both competitions, establishing itself as an industry leader and providing further evidence of PEG's ability to provide accurate, reliable AI scores. These results (available from the links below) demonstrate the viability of AI scoring in general and MI's leadership in particular.

- http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf
- http://www.kaggle.com/c/asap-sas

In addition to the ASAP results, there is a wealth of research that examines the validity and reliability of automated scoring, particularly as it relates to summative assessment, including a large body of work conducted by Dr. Page himself (a representative sample is attached) over nearly 40 years.

Formative Assessments

PEG has also been used to provide tens of millions of scores to students in formative writing

assessments, with over three million essays scored in the last year alone. In addition to providing real-time scores, PEG also adds value when used in a formative context by providing response-specific feedback to the students on the grammar and spelling errors found in their essays, as well as offering targeted instructional feedback on how to improve their writing skills. PEG is in widespread use as an AI scoring engine for formative writing practice websites, including Educational Records Bureau's Writing Practice Program (WPP), Utah State Office of Education's Utah Write, Connecticut State Department of Education's CBAS Write, North Carolina's NC Write, and Learning Express Advantage.

Although less research has been conducted on the efficacy of automated scoring in formative assessment, recent research related to PEG supports the claim that automated scoring technology can be effective in accurately identifying struggling writers in need of specialized interventions beyond practice and revision.[1] MI is pleased to have been selected to participate in the third phase of ASAP research, the Classroom Trials, beginning in Fall 2013, which should serve to further formative-assessment research. The emphasis in the first two phases of ASAP was on evaluating the degree to which current high-stakes writing assessments might be scored through automated methods. The Classroom Trials phase, on the other hand, examines the role that automated scoring might play in helping students achieve higher levels of proficiency in writing through formative assessment (a system that supports frequent evaluation coupled with directed feedback) and in assisting teachers in the design and development of effective individualized instructional strategies.

PEG Functional Description

MI's AI scoring engine is able to automatically score a variety of constructed response items, from multi-page essays to short answers that comprise only a few words, and can work with any number of predefined score-point ranges and rubric definitions. PEG's flexibility allows us to build AI models using the methods that are most effective for each type of response, working equally well on short answers graded for content, essays graded for style, and essays graded for both style and content.

The ability of the engine to match or exceed human reliability depends on a number of factors, including the amount and quality of the training data, the complexity of the item to be scored, and the amount of time available to fine-tune the models. Like most AI scoring engines, PEG relies on an accurate sampling matrix of the anticipated testing population, although there is considerable variability, depending on the complexity of the item, in the number of responses required to build a reliable AI scoring model. While PEG can build models with any amount of training data, we find that a good rule of thumb for achieving high quality models is to provide approximately 200-300 responses per score point, randomly sampled from the testing population. When gathering the training data, we generally require two independent human scores per response. While PEG only requires one score per response to build a model, the second score provides necessary information about how well two humans are able to agree on a score, which is then used as a benchmark for how well PEG's predictions should agree with the human scores.

Model Building

To build a scoring model, PEG analyzes the training set and calculates features that pertain to the content in question. PEG then sends the features to dozens of different algorithms that compete to see which ones can best associate the features with the human-assigned scores. These algorithms draw on many of the latest advances in the field of machine learning to generate both linear and non-linear models. The strongest models are then automatically blended together to create a final model that retains the best elements from the various algorithms. There are six elements of the model-building process:

Representation Generation

Representations are different "views" of each response. For instance, a spell-corrected version of the response might be one representation, while all of the words replaced with a code for their part of speech might be another.

Feature Generation

Those elements of a response that can be measured with a numeric value are referred to as features. We divide the concept of features into two subcategories, mutable and immutable. When a feature can be generated from the response considered in isolation from all other responses, we call it an immutable feature. Things like length of the response, number of grammatical errors, and ratio of the word "the" to other words in the response would all be immutable features. Mutable features are any features that measure information in the response with respect to the set of other responses in the training data. An example of a mutable feature would be deviation from average response length.

Note that the same feature might be measured on many different representations.

Dimensionality Reduction

Due to either the computational complexity or the nature of the algorithm, some of our prediction methods perform poorly when the feature space is massive. In order to reduce that space, we have a number of dimensionality-reduction algorithms that can perform feature selection and/or feature extraction on the data. Feature selection is the process of choosing a subset of features in order to optimize a particular constraint. We generally try to find the features that carry the most information but are the least correlated with one another, although there are other feature selection methods as well. Feature extraction is the process of performing matrix operations on the feature space in order to reduce the size. This creates a new set of implicit features that are functions of the original features.

Learning

Learning is the process of building a model that can predict scores on responses that were not in the original training set. To do this, a learning algorithm uses a set of features for all of the responses in the training data and creates a formula that approximates the scores for each response when given the response's features. Each Learner can also take zero or more hyperparameters, which alter the way the Learner constructs a model. When trying to find the best model, we optimize over a vast space of possible Learner and hyperparameter combinations.

Ensembling

Ensembling is similar to learning, except it generates a formula that maps a collection of learned predictions (from the Learners) to a final prediction. Machine learning research has shown that building a model that blends predictions in this way consistently outperforms any of the individual prediction models.[2] Like Learners, Ensemblers can take zero or more hyperparameters and the space of possible ensemble and hyperparameter combinations is also very large.

Cross-validation

PEG is essentially performing an optimization search across all representations, features, feature selectors, feature extractors, learners (with all of their possible hyperparameters), and ensemblers (with all of their possible hyperparameters) to determine which set of these will most accurately model the scores given to the responses in the training data. One of the risks inherent in machine learning is over-fitting the data. This means that it is possible to home in on particular elements of the responses in your training data in such a way that the model does not generalize well to unseen data. To mitigate this risk, we use a process known as cross-validation, which allows us to develop an estimate for how well any given model will perform on unseen data. This allows our optimization process to identify models that not only perform well on the training data, but will continue to give reliable scores on new responses.

Scoring the NAPLAN Scripts

How the NAPLAN Criteria Were Measured

The model-building process outlined above was used to build separate models for each of the NAPLAN criteria. When we begin the process, we do not know *a priori* what representations, features, feature selectors, feature extractors, learners, or ensemblers will lead to the model that best predicts student scores on a given criterion. PEG's job is to examine all of these available elements and select the ones that best model the scores humans have assigned to the responses in the training data. In order to ensure that PEG is capable of modeling these criteria, the researchers at MI have developed a large set of explicit features for PEG to examine on every response. It is often the case that a given feature will play an important role in multiple criteria. This makes intuitive sense because the criteria are generally highly correlated with one another. That is, a good writer will tend to score well on multiple criteria and a bad writer will tend to score poorly on multiple criteria. The NAPLAN criteria and an overview of PEG's associated feature groups are outlined below.

In addition to the explicit feature groups listed below, PEG measures a theoretically infinite set of implicit features on the responses. These implicit features are based on character, part of speech, and word sequences, recombined in a multitude of ways. When crafting the kind of explicit features listed below, our researchers wrote code that would look for certain predefined patterns in the response and then apply transformations to the patterns to generate a numerical value. Our implicit features, on the other hand, use the data in the training set itself to determine which patterns are highly indicative of performance on the scoring criterion in question. These will be encoded, in various ways, into the saved scoring

models, without us ever seeing them. In many cases it would not be possible to view them explicitly, because they are too complex. We view this model-complexity as a strength. It frees us from the limitations of only looking for features that we happen to think of and instead allows us to leverage the collective intelligence of the humans who scored the responses in the training set. Because of this flexibility, our implicit features tend to perform well on all criteria, be they the style-based criteria found in essay rubrics or the content-based criteria found in short answer and constructed response items.

Moreover, we find that our best performance comes from using both the implicit and explicit features, effectively combining the intelligence of the human readers and our computational linguist researchers. In the following section we attempt to identify those explicit features that appear to align most closely with the constructs being measured by the NAPLAN criteria. Please note that each of the feature groups listed represents a collection of many individual features measured on each response.

Audience

The writer's capacity to orient, engage and persuade the reader.

Associated feature groups:

- Measures of uniqueness
- Sentence structure diversity
- Important sequences of words relevant to scoring the response
- Analysis of word choice including words that may reveal values, attitude, important context, and persuasive techniques
- Detection of informal or colloquial usage
- Detection of audience awareness or the lack thereof through paragraph and response length and specificity of word choice

Text Structure

The organization of the structural components of a persuasive text (intro, body and conclusion) into an appropriate and effective text structure.

Associated feature groups:

- Word sequence analysis which measure lists of key statements, important structure, and reasoning
- Analysis of word choice including words that may identify structural components, evidence, and the writer's position
- Detection of repetitive sentence structure
- Detection of the presence or absence of transitional devices
- Identification of introductions and conclusions
- Detection of unsophisticated attempts at organizing material

- Detection of the expression of personal opinions
- Measures of specificity
- Frequency and types of grammar errors
- Sentence complexity
- Measures of the number, size, and distribution of paragraphs in the response
- Measures of the variability of sentence length within a paragraph

Ideas

The selection, relevance and elaboration of ideas for a persuasive argument.

Associated feature groups:

- Word sequence analysis that measures relations between ideas, support of ideas, and reasoning
- Analysis of word choice including words that identify important ideas, evidence, and persuasive techniques
- Topic modeling that identifies coverage of important ideas and topics
- Detection of specificity and vagueness
- Detection of complex noun and verb phrases
- Standard readability measures
- Sentence complexity
- Frequency of clichés
- Frequency of words that indicate subjectivity

Persuasive Devices

The use of a range of persuasive devices to enhance the writer's position and persuade the reader.

Associated feature groups:

- Word sequence analysis that measures the complexity of persuasive devices and lists of statements
- Analysis of word choice including words that may identify values, evidence, and persuasive techniques
- Measures of sustained use of important words, ideas, and phrases
- Detection of the presence or absence of modal verbs
- Detection of continuous and perfect aspect
- Detection of indicators of personal opinion

- Frequency of various types of modal auxiliaries
- Frequency of various types of conjunctions and other connectives

Vocabulary

The range and precision of contextually appropriate language choices.

Associated feature groups:

- Analysis of word choice and phrasing including both content words and grammatical word classes
- Measures of sustained use of sophisticated vocabulary and effective phrasing
- Measures of vocabulary richness
- Detection of specificity and generality for individual words and phrases
- Detection of complex noun and verb phrases
- Detection of complex modal groups
- Detection of agreement errors between nouns and determiners
- Detection of word choice errors
- Frequencies of words in the response, measured in various ways
- Frequency of words from predetermined word lists that group words of similar difficulty levels
- Measures correlating word length with correct spelling
- Measures of hypernymy and hyponymy by part of speech

Cohesion

The control of multiple threads and relationships across the text, achieved through the use of referring words, ellipsis, text connectives, substitutions and word associations.

Associated feature groups:

- Measures of how the cohesion of a given response matches that of other similar texts
- Analysis of word choice including words that may identify the use of synonyms, related terms, and connectives
- Measures of the use of verb ellipses
- Measures of pronoun usage
- Cohesion as identified by root morphemes
- Morpheme counts
- Measures of the location of main verb and subject within the sentence
- Measures of parenthetical words and phrases

- Measures of transitional words
- Frequency and location of conclusion words

Paragraphing

The segmenting of text into paragraphs that assists the reader to follow the line of argument.

Associated feature groups:

- The number and length of paragraphs
- Detection of the presence of transitional devices between paragraphs
- Measures of the distribution of paragraphs in the response
- Measures of the variability of sentence length within a paragraph

Sentence Structure

The production of grammatically correct, structurally sound and meaningful sentences.

Associated feature groups:

- Sequences of part of speech tags
- Identification of sentence complexity through sentence length and punctuation
- Detection of run-on sentences, comma splices, sentence fragments and excessively long sentences
- Detection of agreement errors and verb tense errors
- Detection of repeated and missing words
- Measures of sentence structure variation within the response
- Measures of the location of relative and other dependent clauses within the sentence or phrase
- Measures of shorter sentences and those with simple structures
- Measures of the location of main verb and subject within the sentence
- Frequency of various types of conjunctions and other connectives

Punctuation

The use of correct and appropriate punctuation to aid reading of the text.

Associated feature groups:

- Detection of the absence or misuse of commas, question marks, hyphens, apostrophes, colons, semi-colons and full stops
- Detection of capitalization errors
- Detection of errors within quotation marks

- Frequencies of punctuation marks by type and total
- Detection of the presence of unbalanced parentheses

Spelling

The accuracy of spelling and the difficulty of the words used.

Associated feature groups:

- Frequencies of spelling errors of varying degree
- Weighted sums of detected spelling errors
- Detection of homophone and other "real word" spelling errors
- Detection of nonstandard spellings
- Frequency of spelling errors correlated to other word characteristics such as length and difficulty
- Comparison between the counts of other word features when misspelled words are omitted vs. when they are corrected
- The frequencies of words in the response, measured in various ways
- The number of words from predetermined word lists that group words of similar difficulty levels
- Measures correlating word length with correct spelling
- Measures of hypernymy and hyponymy by part of speech

Evaluation Metric

When PEG builds a model, it selects the model elements that maximize scoring accuracy for the data in question. Therefore, it is important to choose an agreement statistic on which PEG can optimize its models in such a way that the final model will exhibit reliable, accurate scoring. The inter-rater reliability of two human raters is often measured via perfect/adjacent agreement or the Pearson product-moment correlation coefficient (Pearson's *r*). However, these two metrics each have significant disadvantages. Perfect/adjacent agreement is highly influenced by the overall scale and underlying distribution[3] of the "true" scores, while Pearson's *r* is insensitive to mean difference between raters.[4] We have found that using quadratic weighted kappa, which has become the industry standard for AI scoring, as the optimization and evaluation metric leads to the most reliable and accurate scoring. Quadratic weighted kappa as a metric can detect changes in mean difference and variance between raters and is therefore well suited for comparing the accuracy of AI scoring with respect to human scoring, as well as measuring the agreement of two independent human raters. For the sake of clarity in the discussion below, we refer to quadratic weighted kappa between PEG and Reader 1 as κ_ω(PEG, R1) and quadratic weighted kappa between Reader 1 and Reader 2 as κ_ω(R1, R2).

Even though quadratic weighted kappa performs well as an optimization metric, there are still some deficiencies in using it as an evaluation metric. Quadratic weighted kappa is far less

influenced by the overall scale and underlying distribution of the "true" scores than perfect/adjacent agreement, but it does still display some sensitivity to those aspects of the data. In addition, while AI scoring can outperform human scoring with regard to scoring accuracy, the quality of the human scoring data has a significant impact on PEG's ability to accurately model the data. That is, a low κ_ω(R1, R2) will usually lead to a low κ_ω(PEG, R1). Because of these issues with sensitivity to scale and distribution of scores and being bound by the quality of the training data scores themselves, it is difficult to give a fixed number for what an acceptable value would be for κ_ω(PEG, R1). Instead, we prefer to use the difference between κ_ω(PEG, R1) and κ_ω(R1, R2) as our evaluation metric. We define that value as follows:

$$\Delta_\kappa = \kappa_\omega(\text{PEG, R1}) - \kappa_\omega(\text{R1, R2})$$

When Δ_κ is positive, PEG's scores are more in agreement with Reader 1 than Reader 1's scores are in agreement with Reader 2. When Δ_κ is negative, the opposite is true, Reader 1 and Reader 2 show higher agreement levels than PEG and Reader 1. Of course, in both cases the absolute value of Δ_κ maintains its weight as a relative value between the two kappa values. That is, a larger Δ_κ means more separation between the two kappa values being compared.

Δ_κ is a good metric to quickly show how accurately PEG was able to score a set of data with respect to how accurate human raters are on the same data, but we also report other metrics that our clients may be more familiar with, such as perfect/adjacent agreement, Pearson's *r*, and standard mean difference. However, since PEG was optimized on quadratic weighted kappa, κ_ω and Δ_κ are the best reflection of actual performance. To give some sense of the significance of the values of κ_ω and Δ_κ, Table 1 shows the quadratic weighted kappa and the difference from the rank 1 value for the top 20 competitors of the ASAP Phase 1 and Phase 2 competitions.

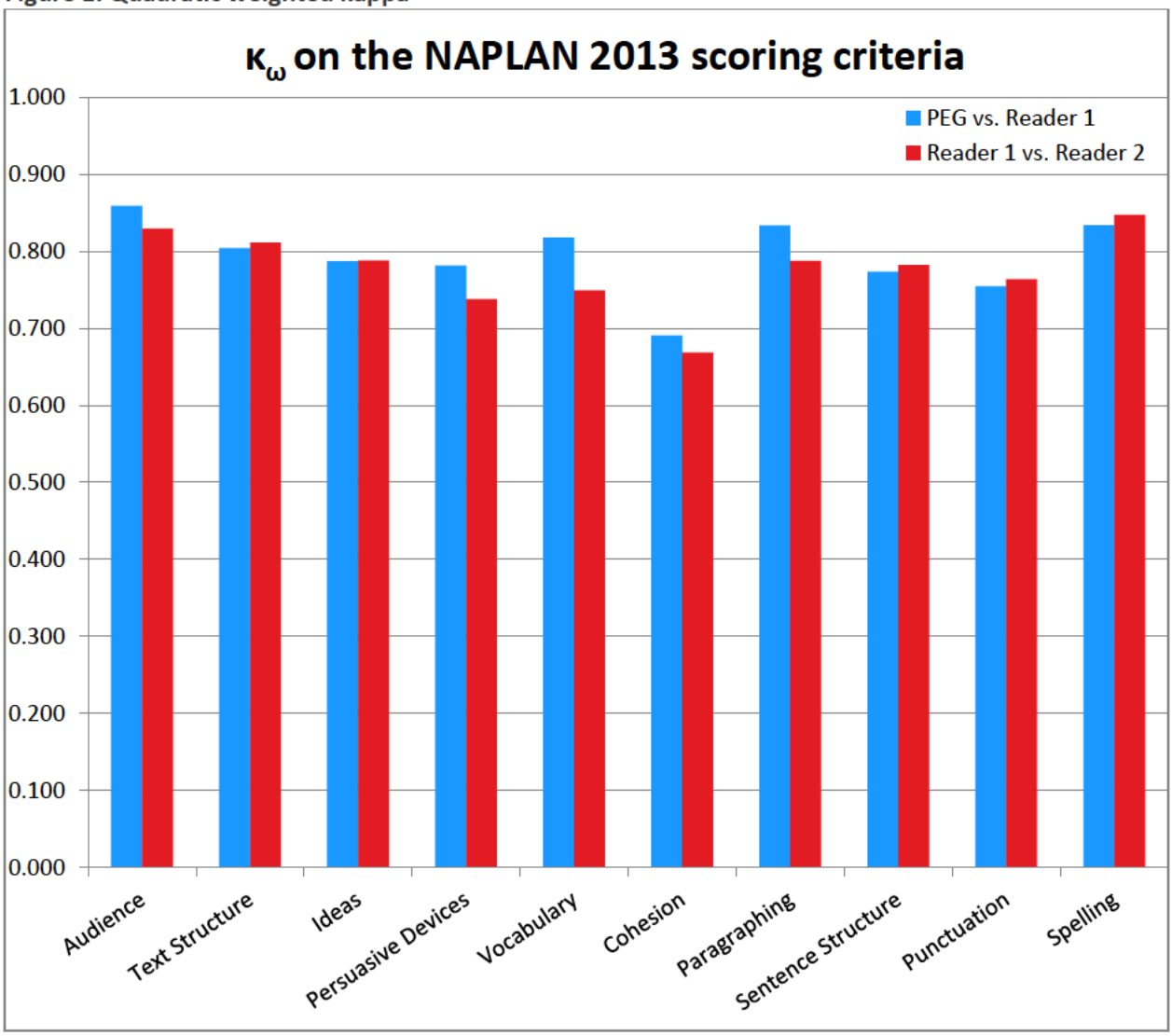
Table 1. The top 20 competitors from the ASAP Phase 1 and Phase 2 competitions

	ASAP Essay		ASAP Short Answer	
Rank	κ_ω	Δ_κ	κ_ω	Δ_κ
1	0.814	0.000	0.748	0.000
2	0.809	-0.005	0.747	-0.001
3	0.806	-0.008	0.739	-0.009
4	0.804	-0.010	0.737	-0.011
5	0.799	-0.015	0.735	-0.013
6	0.797	-0.017	0.735	-0.013
7	0.788	-0.026	0.734	-0.014
8	0.788	-0.026	0.731	-0.017
9	0.786	-0.028	0.726	-0.022
10	0.784	-0.030	0.725	-0.023
11	0.773	-0.041	0.721	-0.026
12	0.772	-0.042	0.716	-0.032
13	0.765	-0.049	0.716	-0.032
14	0.764	-0.050	0.712	-0.036
15	0.762	-0.052	0.707	-0.041
16	0.762	-0.052	0.706	-0.042
17	0.755	-0.060	0.705	-0.043
18	0.754	-0.060	0.704	-0.044
19	0.753	-0.062	0.703	-0.045
20	0.752	-0.062	0.701	-0.047

Analysis of ACARA 2013 Results

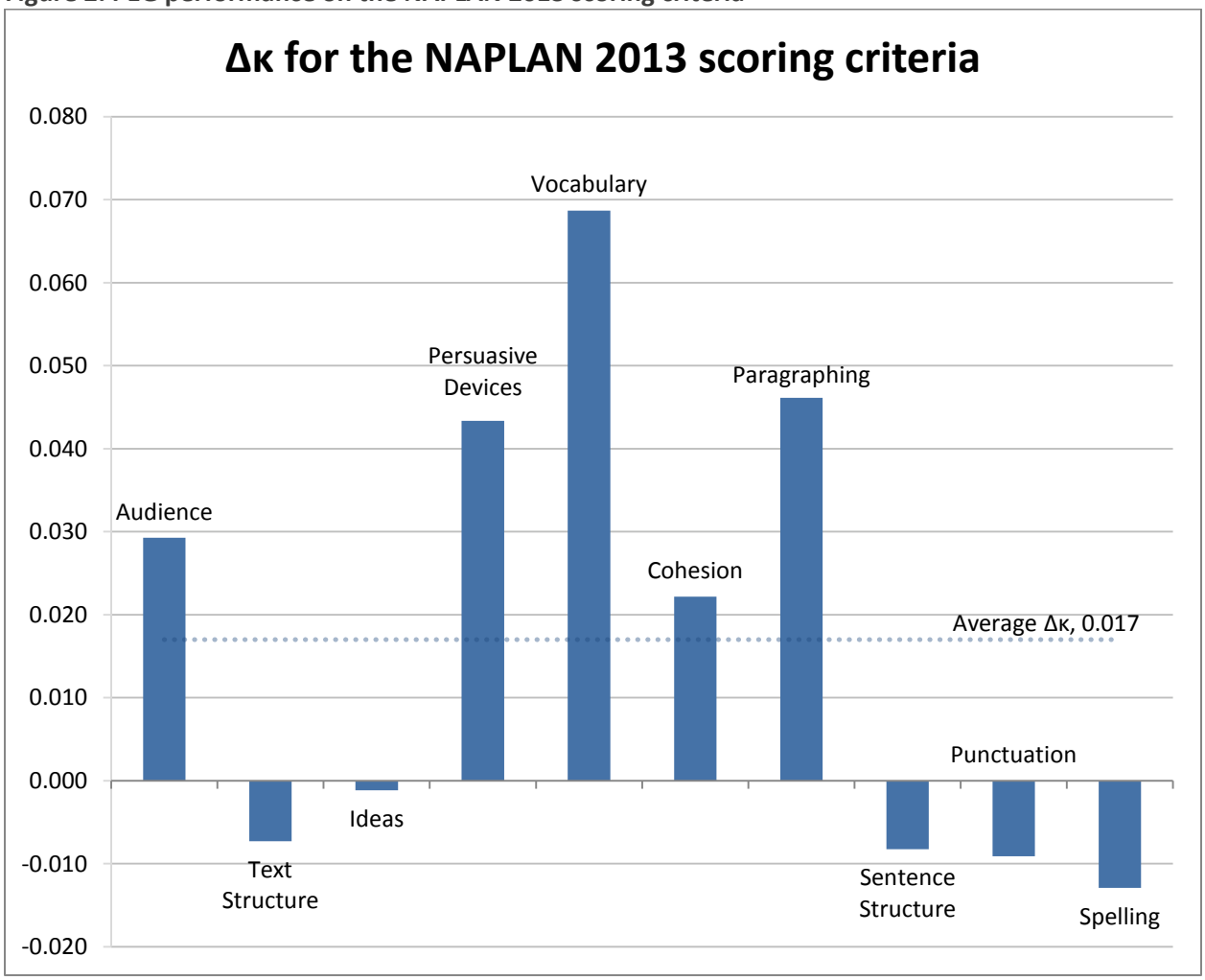
The ACARA human readers set a very high bar in terms of inter-rater agreement. High agreement between two independent human readers means that the scoring was generally accurate and precise. Because of this, PEG was able to generate predictions that matched the human scores very closely. As you can see in Figure 1, the quadratic weighted kappa for each criterion wound up in a narrow band near the top of the scale.

Figure 1. Quadratic weighted kappa



Our analysis, however, shows that PEG's agreement with the human scores was generally better than the human inter-rater agreement, the yardstick by which we measure PEG's performance. Although the sign of Δ_κ was split evenly across the criteria, the average Δ_κ across the criteria was positive. Figure 2 shows the Δ_κ between PEG vs. Reader 1 and Reader 1 vs. Reader 2 for the 10 criteria. The dotted line represents the average of all 10 criteria. Note that even when the sign of Δ_κ was negative, the quadratic weighted kappas were so close that we would consider our model accuracy to be equivalent to that of humans.

Figure 2. PEG performance on the NAPLAN 2013 scoring criteria



Tables 2 and 3 display these statistics for each criterion as well as Pearson's *r*, perfect/adjacent agreement, and standardized mean difference.[5]

Table 2. Agreement statistics for NAPLAN criteria 1-5

	Audience	Text Structure	Ideas	Persuasive Devices	Vocabulary
QWK					
PEG	0.859	0.804	0.787	0.781	0.818
Human	0.829	0.811	0.788	0.738	0.749
Delta	0.029	-0.007	-0.001	0.043	0.069
Pearson's r					
PEG	0.862	0.812	0.791	0.783	0.822
Human	0.833	0.813	0.790	0.741	0.758
Delta	0.030	-0.001	0.001	0.042	0.065
Perfect					
PEG	0.640	0.652	0.628	0.605	0.676
Human	0.643	0.684	0.655	0.602	0.643
Delta	-0.003	-0.032	-0.027	0.003	0.032
Adjacent					
PEG	0.982	0.991	0.976	0.997	0.988
Human	0.976	0.997	0.985	0.979	0.976
Delta	0.006	-0.006	-0.009	0.018	0.012
Standardized Mean Difference					
	0.010	0.041	0.012	0.046	0.017

Table 3. Agreement statistics for NAPLAN criteria 6-10

	Cohesion	Paragraphing	Sentence Structure	Punctuation	Spelling
QWK					
PEG	0.691	0.834	0.774	0.755	0.834
Human	0.668	0.787	0.782	0.764	0.847
Delta	0.022	0.046	-0.008	-0.009	-0.013
Pearson's r					
PEG	0.692	0.836	0.781	0.765	0.838
Human	0.674	0.789	0.787	0.765	0.847
Delta	0.018	0.047	-0.006	0.001	-0.010
Perfect					
PEG	0.699	0.667	0.534	0.510	0.640
Human	0.690	0.643	0.593	0.640	0.687
Delta	0.009	0.024	-0.059	-0.130	-0.047
Adjacent					
PEG	0.991	0.988	0.968	0.976	0.991
Human	0.988	0.979	0.979	0.976	0.979
Delta	0.003	0.009	-0.012	0.000	0.012
Standardized Mean Difference					
	0.009	0.002	0.041	0.008	0.023

Finally, Table 4 shows the same statistics for totals and averages. TOT represents the total score given to the student. This score is generated by summing the ten NAPLAN criteria, as specified in the ACARA contract. AVG shows the average of the statistics across the ten NAPLAN criteria.

Table 4. Totals and averages of agreement statistics for NAPLAN criteria

	TOT	AVG
QWK		
PEG	0.912	0.793
Human	0.917	0.776
Delta	-0.004	0.017
Pearson's r		
PEG	0.919	0.798
Human	0.920	0.780
Delta	-0.001	0.019
Perfect		
PEG	0.112	0.625
Human	0.153	0.648
Delta	-0.041	-0.023
Adjacent		
PEG	0.354	0.985
Human	0.419	0.982
Delta	-0.065	0.003
Standardized Mean Difference		
	0.016	0.021

Observations

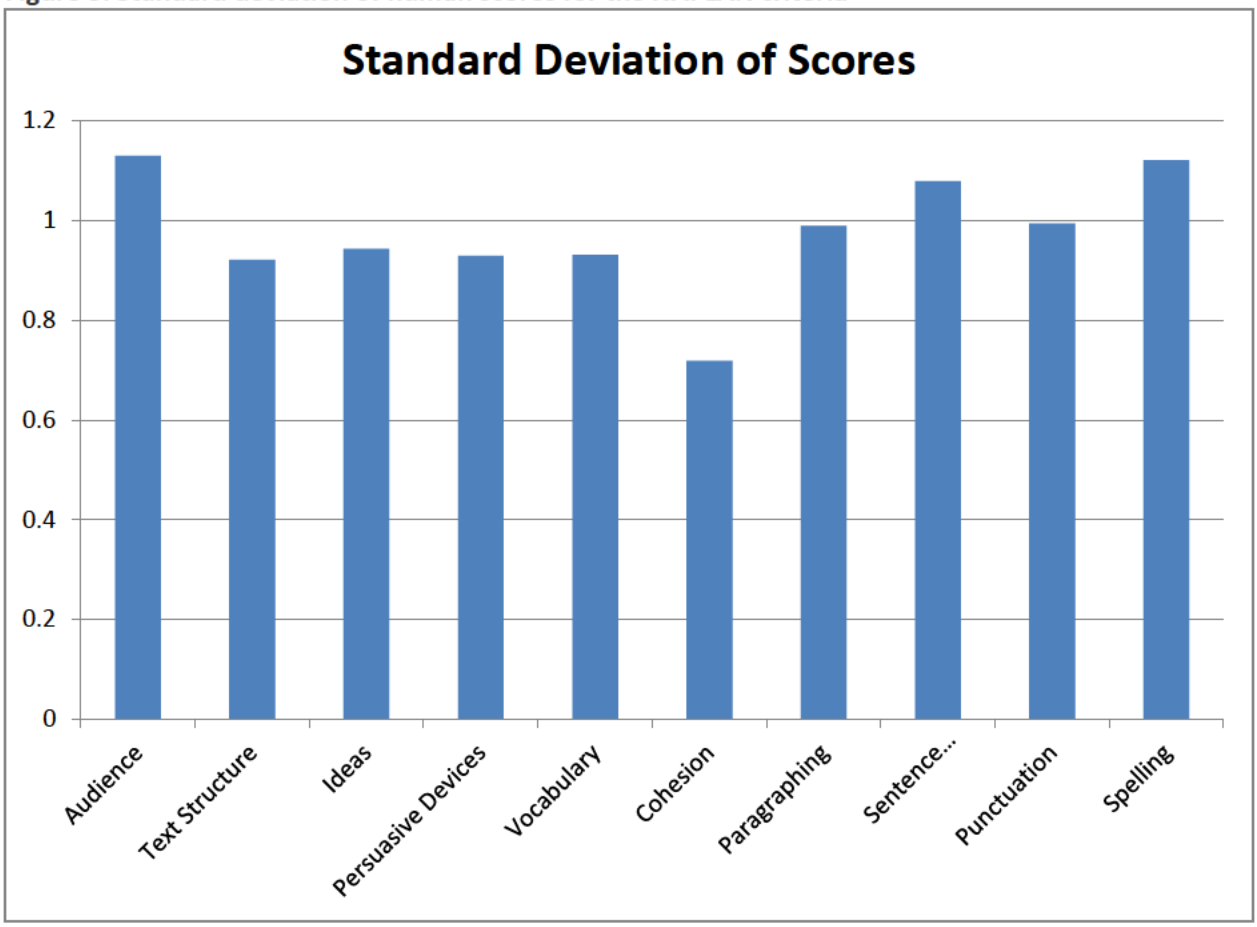
As is clear from the analysis above, MI's AI scoring engine, PEG, demonstrated quite clearly that it could provide scores that were equal to and, in some cases, better than those provided by human readers. The differences between human reliability and the AI scoring reliability were often small and generally favored PEG. In what follows, we provide some additional ad hoc analyses of the outcomes from the ACARA 2013 automated scoring study that we feel are worthy of note.

Cohesion

Generally, when there are quadratic weighted kappa values of less than 0.70, further study is required from our data analysts to ensure that there is not something wrong with the model that PEG has created. In the ACARA 2013 study, the only criterion to fall below that cut-off was "Cohesion" with a quadratic weighted kappa of κ_ω(PEG, R1) = 0.691. However, on this criterion, the human-human kappa, κ_ω(R1, R2), was also below the 0.70 threshold, with a value of κ_ω(R1, R2) = 0.668. When human agreement is low, it is difficult for PEG to build a model with agreement that is substantially higher. In this case, since PEG's score is higher than the human score, and they are both relatively close to the threshold, we feel reasonably confident that it is a good model.

An interesting thing to note is that, while "Cohesion" had the lowest human-human quadratic weighted kappa of the criteria, it also had the highest human-human perfect agreement. To understand why κ_ω would be low when the perfect agreement was so high, we examined the underlying data. What we found was that this criterion was an outlier in terms of standard deviation. In particular, the standard deviation of the human scores was lower than any other criterion, as you can see in Figure 3.

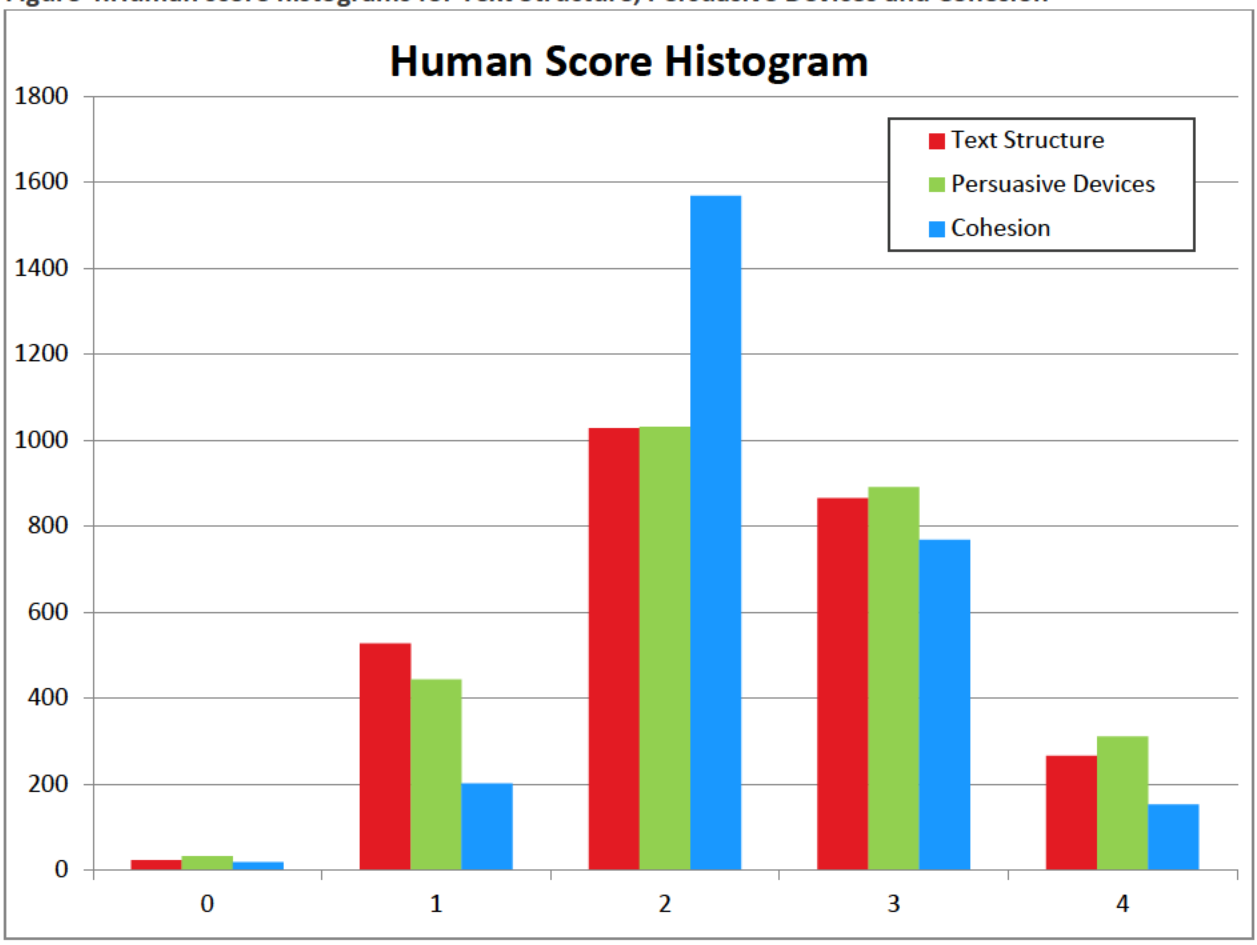
Figure 3. Standard deviation of human scores for the NAPLAN criteria



It is particularly informative to view this criterion in comparison with the other two criteria that have a 0-4 score range, "Text Structure" and "Persuasive Devices." You can see from the

histogram in Figure 4 that, while they are all normally distributed, "Cohesion" has a higher peak with a larger positive skew. This may be a potential area of review for NAPLAN to improve the rubric.

Figure 4.Human score histograms for Text Structure, Persuasive Devices and Cohesion



Punctuation

Another criterion that raised a potential flag for us was "Punctuation." On this criterion, PEG's quadratic weighted kappa was slightly lower than the human-human value, with Δ_{κ} = -0.009. This is well within the bounds of acceptability for a model, especially considering κ_{ω}(PEG, R1) was 0.755. What concerns us is that the perfect agreement delta on this criterion is abnormally low, with a value of -0.130. To understand more about this criterion, we calculated the correlation coefficient for all of the criteria, as shown below. Table 5 is a heat map of the correlation of human scores between criteria. You can see that Punctuation is the least correlated criterion with respect to scoring on the other criteria. While this does not directly explain the aforementioned perfect agreement delta, it may point to some irregularity in the scoring that is being reflected in the agreement statistics.

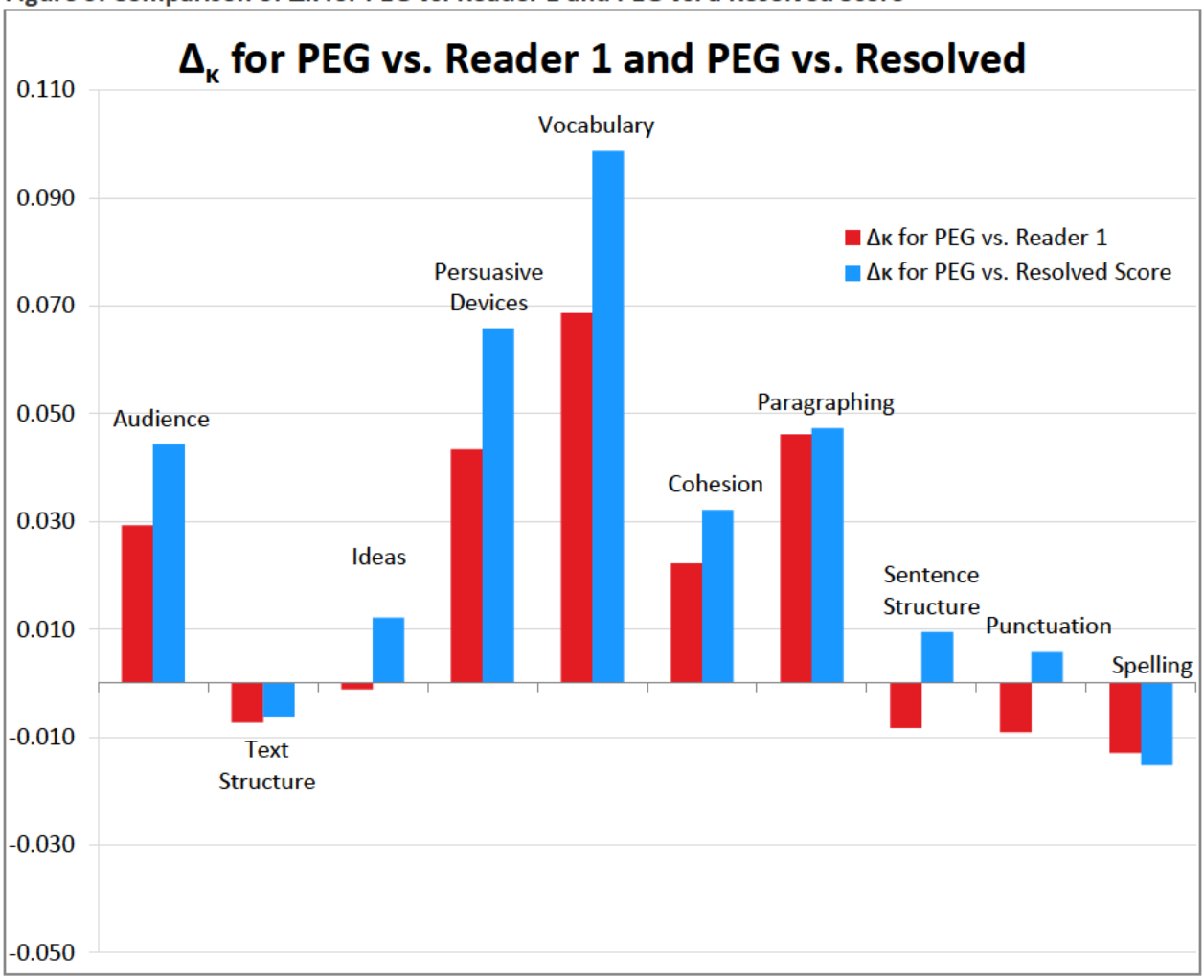
Table 5. Human score correlation coefficient heat map across NAPLAN criteria

	Audience	Text Structure	Ideas	Persuasive Devices	Vocabulary	Cohesion	Paragraphing	Sentence Structure	Punctuation	Spelling
Audience	1	0.8491	0.8535	0.8516	0.8416	0.794	0.8093	0.8396	0.7033	0.8305
Text Structure	0.8491	1	0.785	0.8071	0.7698	0.7441	0.8116	0.7958	0.6798	0.7779
Ideas	0.8535	0.785	1	0.799	0.7946	0.7473	0.7557	0.7929	0.6502	0.7942
Persuasive Devices	0.8516	0.8071	0.799	1	0.8003	0.7461	0.7929	0.7755	0.6359	0.7815
Vocabulary	0.8416	0.7698	0.7946	0.8003	1	0.7668	0.7471	0.7847	0.6286	0.7612
Cohesion	0.794	0.7441	0.7473	0.7461	0.7668	1	0.7326	0.7753	0.6571	0.7057
Paragraphing	0.8093	0.8116	0.7557	0.7929	0.7471	0.7326	1	0.7677	0.6331	0.7569
Sentence Structure	0.8396	0.7958	0.7929	0.7755	0.7847	0.7753	0.7677	1	0.7371	0.7882
Punctuation	0.7033	0.6798	0.6502	0.6359	0.6286	0.6571	0.6331	0.7371	1	0.6823
Spelling	0.8305	0.7779	0.7942	0.7815	0.7612	0.7057	0.7569	0.7882	0.6823	1

Resolved Score

This report has repeatedly referred to PEG's level of agreement with Reader 1. We assume, as is usually the case, that the idea of Reader 1 is a convenience and that the first set of human scores is actually generated by a number of human readers. One of the benefits of AI scoring, in addition to the obvious cost and speed advantages, is the potential to smooth out inconsistencies among individual readers or even a lack of consistency within the scores of a single human reader. One way to see this is to examine PEG's agreement levels with a resolved score. If we posit that taking the average of the two independent human scores brings us closer to the "true" score for each response, we can calculate PEG's quadratic weighted kappa with that new score and compare it to the quadratic weighted kappa we measured between PEG and Reader 1. Figure 5 shows the Δ_κ values for the original $\Delta_\kappa = \kappa_\omega(\text{PEG, R1}) - \kappa_\omega(\text{R1, R2})$ and a resolved $\Delta_\kappa = \kappa_\omega(\text{PEG, Resolved}) - \kappa_\omega(\text{R1, R2})$. Note that on almost every criterion, comparing the PEG scores to a resolved score increases the Δ_κ, sometimes even changing the sign.

Figure 5. Comparison of Δκ for PEG vs. Reader 1 and PEG vs. a Resolved Score



We do not present this as evidence of PEG's accuracy because these models were built from the scores of Reader 1, and only a comparison with Reader 1 offers an apples-to-apples evaluation. Instead, we offer this as a potential improvement to the process. By adding a resolved score and using that data to build our models, we would expect to be able to predict scores on these criteria with even more accuracy.

Data Anomalies

As a final note, we did find a few anomalies in the data, which was otherwise very clean. The first, which we mentioned in earlier email communications with ACARA, is that in the released scores for the test set, the total score for Reader 1 (S1TOT), as well as the final score (Save) and the total difference (SDif), were miscalculated. For all of those values, the sample number (Sample) was mistakenly included in the total score for Reader 1 (S1TOT) when the sample was 3 (the test set). The second anomaly, which we didn't find until after the model building and predictions had completed, was that on script #1326080 in the training data, the Ideas criterion was scored as a 9 by Reader 1, which is outside the rubric range of 0-5 for that criterion. Neither of these issues created any problems for our AI scoring process; we were able to adjust the scores for the first issue before our analysis, and the second issue was treated as data noise

by our model building process. However, we mention these issues here in case either of them represent something more than mere transcription errors in the ACARA scoring process.

Conclusion

MI, long recognized as an industry leader in the handscoring of student essays and, more recently, as a leader in the automated scoring of essays, partnered with ACARA to test the feasibility of using AI scoring on the NAPLAN Writing Test. Specifically, MI deployed its award-winning AI engine, PEG, to build models and predict scores for the ten NAPLAN criteria, using the handscored student essays provided by ACARA.

For each of the ten criteria that make up the overall scoring rubric for the provided essays, PEG used technologies that draw on some of the latest advances in artificial intelligence and natural language processing to identify features across multiple levels of analysis. Those features were then fed into a system that optimized models to predict scores. By all measures – quadratic weighted kappa, Pearson's *r*, perfect and adjacent agreement – PEG predicted scores that were equivalent to, and arguably better than, those provided through human scoring. Results were further enhanced when resolved scores – an average of the scores provided by the two human readers – were used as the standard against which PEG's predicted scores were compared.

These results provide support for the feasibility of using AI scoring in a large-scale, ongoing writing program to provide scores that represent a valid assessment of student performance.

References

Calkins, L. (2006). *A Guide to The Writing Workshop, Grades 3-5*. Portsmouth, NH: First Hand.

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An investigation of the impact of the 6+1 Trait Writing model on grade 5 student writing achievement* (NCEE 2012–4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Council of Chief State School Officers. (2008). *Attributes of Effective Formative Assessment*. FAST SCASS Assessment Subcommittee [Electronic Version]. Washington, DC: Council of Chief State School Officers.

Culham, R. (2003). *6 + 1 Traits of Writing: The Complete Guide*. Scholastic, Inc. New York, New York.

Graham, S., Harris, K., & Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment.* A Carnegie Corporation Time to Act report. Washington, DC: Alliance for Excellent Education.

Shermis, M. D., & Hammer, B. (2012). Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. Retrieved May 3, 2013, from ASAP: http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Wiggins, G. & McTighe, J. (2001). What is Backward Design? In *Understanding by Design* (1st ed.) (pp. 7-19). Upper Saddle River, NJ: Merrill Prentice Hall.

Wilson, J., & Andrada, G. (2013, April). *Examining Patterns of Writing Performance of Struggling Writers on a Statewide Classroom Benchmark Writing Assessment: The Utility of Dynamic Assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Selected PEG Research

Hiller, J. H. (1998). *Applying computerized text measurement strategies from Project Essay Grade (PEG) to military and civilian organizational needs.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, ED418995.

Page, E. B. (1966). *The imminence of grading essays by computer*. Phi Delta Kappan, v. 48, pp. 238-243.

Page, E. B. (1967a). *Grading essays by computer: progress report*. Proceedings of the 1966 Invitational Conference on Testing. Princeton. N J: Educational Testing Service.

Page, E. B. (1967b). Statistical and linguistic strategies in the computer grading of essays. Proceedings of the Second International Conference on Computational Linguistics. 34.

Page, E. B., Fisher. G. A., & Fisher, M.A. (1968). "Project Essay Grade: A FORTRAN program for statistical analysis of prose." *British Journal of Mathematical and Statistical Psychology, 21*, 139.

Page, E. B., & Paulus, D.H. (1968), *The analysis of essays by computer* (Final Report to the Bureau of Research at the U.S. Office of Education for Project 6-1318). Washington. DC: U.S. Department of Health, Education, and Welfare.

Page. E. B. (1985). *Computer grading of student essays*. In T. Husen & T. N. Postlethwaite (Eds.), International Encyclopedia of Educational Research (pp. 944-946). Oxford. England: Pergamon.

Page, E. B. (1994, Winter). "Computer Grading of Student Prose, Using Modern Concepts and Software." *Journal of Experimental Education* 62(2) pp. 127-42.

Page, E. B., Truman, D. L. & Lavoie, M. J. (1994, November 11). *"Teacher's Helper": Proposed use of Project Essay Grade for the English classroom.* Symposium conducted at the annual

meeting of Annual Meeting of the South Atlantic Modern Language Association, Baltimore, MD.

Page, E. & Petersen, N. S. (1995, March). "The Computer Moves into Essay Grading: Updating the Ancient Test." *Phi Delta Kappan (76)*7, pp. 561-65.

Page, E. B., Keith, T. Z., & Lavoie, M. J. (1995, August 13). *Construct Validity in the Computer Grading of Essays*. Handout at the Annual Meeting of the American Psychological Association (APA), New York, NY.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997, March 27). *Computer Analysis of Student Essays: Finding Trait Differences in the Student Profile.* Symposium on Grading Essays by Computer conducted at AERA/NCME, Chicago, IL.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001, June). "On-Line Grading of Student Essays: PEG Goes on the World Wide Web." *Assessment & Evaluation in Higher Education 26*(3), pp. 247-59.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z. & Harrington, S. (2002, February). "Trait Ratings for Automated Essay Grading*." Educational and Psychological Measurement 62*(1), pp. 5-18.

Shermis, M. D., Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). "Trait ratings for automated essay grading." *Educational and Psychological Measurement 62*(1), pp. 5-18.

---

[1] Wilson, J., & Andrada, G. (2013, April). *Examining Patterns of Writing Performance of Struggling Writers on a Statewide Classroom Benchmark Writing Assessment: The Utility of Dynamic Assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

[2] Rokach, L. (2010). "Ensemble-based classifiers." *Artificial Intelligence Review 33*(1-2), 1-39.

[3] Williamson, D. M., Xi, X., & Breyer, F. J. (2012). "A framework for evaluation and use of automated scoring." *Educational Measurement: Issues and Practice 31*(1), 2-13.

[4] Schuster, C. (2004). "A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales." *Educational and Psychological Measurement 64*(2), 243-253.

[5] Williamson, et al. (2012).

Automated Essay Scoring of Writing Scripts: Final Report

Submitted by Pearson Knowledge Technologies

August 30th

Contact: Peter Foltz, Peter.Foltz@pearson.com

Executive Summary

The Australian Curriculum, Assessment and Reporting Authority (ACARA) annually undertakes the responsibility of assessing over one million students in Reading, Writing, Language Conventions, and Numeracy, all through the National Assessment Program—Literacy and Numeracy (NAPLAN). The essays collected from the September-December 2012 Online Writing Pilot Study are being used to investigate the costs and efficacy of automated scoring systems, as compared to human markers. ACARA wishes to engage a contractor.

The Knowledge Technologies (KT) group of Pearson was engaged as a contractor to score the scripts and provide a report on the performance of automated scoring, provided by a trusted resource in the assessment sector. Pearson applied its automated scoring writing technology to the scripts by developing individual scoring models for each of the ten NAPLAN scoring criteria. The system was trained on the first sample of 677 scripts. The ten scoring models were then applied to score a second sample of 340 scripts. The results of the automated scoring of these scripts were compared against the performance of two independent human raters.

The overall results indicate that the automated scoring models could provide scores that agree highly with the individual human raters as well as their averaged scores, as measured by correlation and exact and adjacent agreement. In addition, the models were applied to the third sample of 339 scripts, which KT scored without knowledge of the human scores and the performance of the automated scoring were independently evaluated. Results presented in this report describe the performance across the different scoring criteria as well as by individual score point. Implications are discussed in terms of alignment of the automated scoring approach to the scoring criteria as well as for potential additional refinements.

Overview of Automated Essay Scoring

New assessments are incorporating more items that require students to demonstrate their problem solving skills on authentic, complex tasks in language arts, mathematics, and other content areas. The use of constructed-response (CR) items is growing, increasing reliance on human scoring, intelligent computer scoring or a combination of both. Automated scoring technology is advancing allowing its application to large-scale assessments

Automated scoring systems provide consistency over time and location, which promotes equity, enables accurate trend analysis, and provides comparable results for use at the classroom, school, district, or state level. Automated scoring of CR items has grown rapidly in large scale testing because systems can produce scores more reliably and quickly and at a lower cost than human scoring (see Topol, Olson, & Roeber, 2011). There are several automated systems (at Pearson and elsewhere) able to score CR items, including essays, spoken responses, short text answers to content questions, and numeric and graphic responses to math questions.

In the late 1980s and 1990s, the group now known as the Knowledge Technologies group of Pearson invented many of the key techniques that enable automated scoring of constructed language in assessment tasks. Many of these technologies were initially developed for other applications, including recommendation engines (Foltz & Dumais, 1992; Hill et al., 1995), information retrieval (Deerwester et al., 1989; Streeter & Lochbaum, 1988), machine learning (Landauer & Dumais, 1998), and telephone speech recognition (Bernstein, et al., 1994). In the succeeding 15 years, Pearson has assembled these researchers into an advanced development group where they have focused on research and development of technologies for the assessment field.

The artificial intelligence methods that have been incorporated into existing automated scoring technology include state-of-the-art methods in NLP, large-scale corpus-based analyses, knowledge representation, machine learning, and speech recognition. Development of automated scoring technologies requires understanding how the technologies can be implemented, how they can be combined and incorporated into scoring systems, and how to measure the psychometric effects of applying the methods. For example, there are multiple automated scoring methods that can be applied to scoring items (e.g., Hearst, 2000). The methods often provide close to similar results, making it hard to distinguish them from each other. For example, regression-based hill climbing, neural network models, and Bayesian approaches often provide very similar effectiveness in predicting overall student scores from features extracted from essays.

Additionally, features such as the length of essays are highly co-linear with measures that provide deeper analyses of the quality of content. Thus, as new automated scoring methods are developed, it continues to be important to assess how they improve performance in assessment while accounting for features relevant to the assessment constructs in terms of:

- Improving accuracy of scores compared to existing automated scoring methods
- Improving reliability when compared against human scorers, and validity against external measures

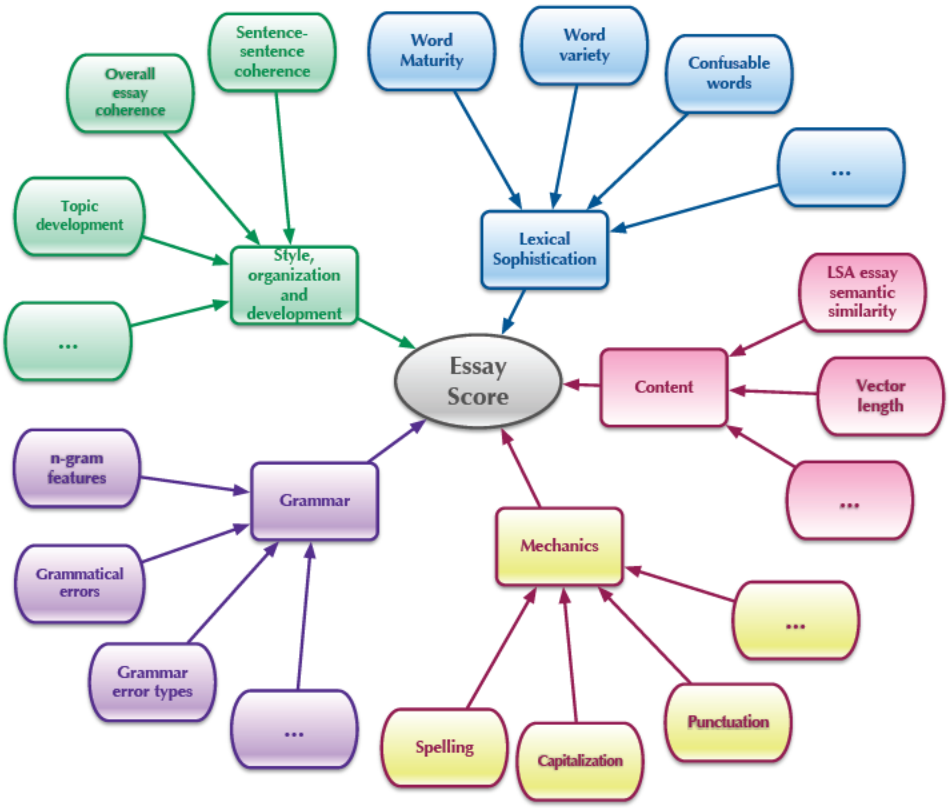
- Addressing additional types of assessment items or feedback that cannot currently be automatically assessed
- Improving the speed or efficiency of returning scores and feedback

In terms of emerging technologies, it is important to examine both where advancements are being made in automated scoring technologies, as well as the areas within assessments that would most benefit from automated scoring technologies.

Generalized approach to automated scoring

The quality of a student's essay can be characterized by a range of features that measure the student's expression and organization of words and sentences; the student's knowledge of the content of the domain; the quality of the student's reasons; and the student's skills in language use, grammar, and the mechanics of writing. The Intelligent Essay Assessor (IEA) evaluates these features using automatically computed measures, including Latent Semantic Analysis (LSA) for content and natural language processing (NLP) techniques to analyze linguistic features of the writing. These features include flow, coherence, word usage, and grammatical constructions, as well as mechanical aspects such as spelling and punctuation. The computational measures extract aspects of student performance that are relevant to the constructs for the competencies of interest (Foltz et al., 2013, Hearst et al, 2000, Williamson et al., 2010). For example, a measure of the type and quality of words used by a student provides an effective and valid measure of a student's lexical sophistication.

Because a student's performance on an essay typically requires showing combined skills across language expression and knowledge, it is critical that the scoring features used in the analysis cover the construct of writing that is being scored. Thus, multiple language features are typically measured and combined to provide a score. The following figure illustrates some of the features used in IEA and how they relate to specific constructs of student writing performance.



Essay Scoring features. IEA evaluates writing using a combination of features that measure aspects of the content, lexical sophistication, grammar, mechanics, style, organization, and development within essays.

IEA combines background knowledge about English in general, and the subject area of the assessment in particular, along with prompt-specific algorithms to learn how to match student responses to human scores. Using a representative sample of responses that are double-scored by humans, the computer compares the content and relevant qualities of the writing of each student response, along with the scores given to the responses by the human scorers. From these comparisons, a prompt-specific algorithm is derived to predict the scores that the same scorers would assign to new responses.

Pearson's Intelligent Essay Assessor (IEA) evaluates the structure, style, and content of writing using a range of AI-based technologies. One key differentiating technology is Pearson Knowledge Technologies' (PKT) unique implementation of Latent Semantic Analysis (LSA), an approach that generates semantic similarity of words and passages by analyzing large bodies of relevant text. LSA can then "understand" the meaning of text, much the same as a human reader. IEA provides an immediate overall evaluation of a response, as well as scores on specific traits. It can be tuned to understand and evaluate text in any subject area. For the present project, the IEA is well suited to be applied to score 1,600 NAPLAN Online Writing Pilot Study scripts.

Evaluating Essay Content

For many kinds of essays, a critical feature for predicting human scores is an essay's content. IEA measures content using Latent Semantic Analysis (LSA), a statistical semantic model invented by principals of the Knowledge Technologies group of Pearson in the late 1980s (Deerwester, et al., 1989;

Landauer & Dumais, 1988). LSA is now in wide use around the world in many applications in many languages, including Internet search, psychological diagnosis, signals intelligence, educational and occupational assessment, and intelligent tutoring systems, as well as in basic studies of collaborative communication and problem solving.

LSA's ability to gauge the quality of a text's meaning at the level of human raters has produced a cottage industry over the last 25 years of new applications where content coverage and quality are the core metrics (e.g., there are about 20,000 references to LSA according to Google™ Scholar). The method has been incorporated into such products as Google's search engine and Apple®'s spam filtering. LSA has also been widely researched as a computational psychological model of the representation of knowledge, including modeling language acquisition, metaphor comprehension, and semantic priming (Kintsch, 2000; Landauer et al., 2006).

LSA derives semantic models of English (and other languages) from an analysis of large volumes of text equivalent to the reading a student may have done through high school (about 12 million words). LSA builds a co-occurrence matrix of words and word frequency in paragraph-sized units and then reduces the matrix by singular value decomposition (SVD), a matrix algebra technique similar to factor analysis. From this analysis, LSA derives a representation of the meaning of words, sentences, paragraphs, and larger units of texts. 300 independent vectors are usually used to represent the meaning of each word and each paragraph in the text collection.

The accuracy of the LSA meaning representation can be assessed by machine-human correlations in rating the similarity of meaning between pairs of paragraphs and the similarity of meaning between pairs of words. Evidence confirms that LSA rates the similarity of meaning between texts about 90 percent (90%) as well as two human scorers would do (Landauer Foltz & Laham, 1998).

For scoring essays, the training-set essays are each given a 300-dimensional score by averaging the word vectors occurring in each essay. That is, each word is represented by a vector with 300 real numbers corresponding to each of the dimensions—the separately measured quantities describing the essay. New essays to be graded are given a 300-dimensional representation based on the words that occur in them and averaged over each of the 300 dimensions. The new essay is then compared to each of the training-set essays in terms of similarity (cosine of the angle between the two essays). The closest neighbors to the new essay and training essays help to determine the content score. Essays with high scores will tend to cluster. So, a new essay close to high scoring training essays will receive a high score. Off-topic essays can be flagged automatically because they have insufficient content similarity to the training papers.

IEA Language Features

Along with content-based measures, a range of other automatically computed measures are also used to score the lexical sophistication, grammatical, mechanical, stylistic, and organizational aspects of essays. The separate characteristics of student essays on which teachers base grades, comments, and corrections influence IEA scores to approximately the same extent they do human scorers.

This is also true of the characteristics described in the rubrics that human scorers seek to follow. Measures of lexical sophistication include measuring the developmental maturity of the words used

(Landauer, Kireyev & Panaccione, 2011), as well as the variety of types of words used. Grammar and mechanics measures use natural language processing (NLP)-based approaches to analyze specific linguistic features of the writing. For grammar, such measures detect run-on sentences, subject-verb agreement, sentence fragments, and use of possessives, among others. For assessing mechanics, measures are used that examine appropriate spelling (including different variants of English), punctuation, and capitalization.

The assessment of stylistic and organizational aspects of essays are evaluated using a combination of LSA-based measures to analyze coherence in the essay, as well as NLP-based measures that assess aspects of the organization, flow and development across the essays. For specific essay types, additional features are incorporated which assess aspects of topic development, such as the strength of an introduction, use of supporting arguments, and the quality of the conclusion. Unless explicitly called for by a test design and documented for users, measures based on raw counts of words, sentences or paragraphs are not included (e.g., counting words, adjectives, number of occurrences of "therefore"). While these measures can be predictive, students can be too easily coached to exploit such count-based measures.

Training the IEA

The Intelligent Essay Assessor uses a machine-learning approach in which it is trained to score based on the collective wisdom of trained human scorers. Based on a sample of human scored student responses, IEA learns the different features that human scorers evaluate when scoring a response, and how the scorers weigh and combine those features to produce a score. IEA is trained based on the scores assigned by human scorers to several hundred representative student responses written in response to a particular prompt for a particular grade level. By using computational modeling, IEA mimics the way humans score.

We typically train IEA using averaged or consensus human scores assigned to each response. Training on average or consensus scores gives IEA a more accurate measure of the true quality of the essays and a more complete picture of the score range by explicitly recognizing responses that are on the cusp between two score points—say, a high 2 or a low 3. By training on the average, we also avoid building in idiosyncrasies related to the scoring rules used to determine a final score. For example, if the scoring rules are such that the final score is the higher of two human adjacent scores, then by training on the final score, IEA may tend to score high compared to a single human scorer.

The responses used to train the scoring engine determine how future responses are scored. It is therefore critical that the sample of student responses used for training and evaluating the scoring engine should represent the full range of student responses and scores. If certain score points are missing or under-represented in the data, it will be difficult for the scoring engine to accurately assign those scores. The responses used to train the system should be 100 percent double-scored by human scorers and also receive resolution scores for non-adjacent agreement. The goal is to have as much, and as accurate, information as possible about how a response should be evaluated.

Responses that are used to train human scorers (e.g., anchor papers) can also be used to train the automated scoring engine, but those responses represent a very small portion of the overall population

and so are not sufficient. Specially selecting particular examples is also usually not helpful. While they can be included, it is best to take a random sample so that a wide range of responses and cases of scorer agreement and disagreement are included. We generally recommend a stratified random sample to verify that a sufficient number of examples at the endpoints are included.

Scoring the NAPLAN criteria (traits)

Human scorers are able to score essays for different traits within essays by focusing on different features of the essay or criteria in their evaluation. For example, to score an essay on conventions, a human scorer would focus on a student's grammar, spelling, and punctuation. Similarly, IEA can generalize to scoring different traits by choosing and weighting different combinations of features. A subset of the features can be used in the training, such as simply choosing features related to conventions if scoring a conventions trait. By then training IEA on human scores, it learns to associate the features within the IEA set that best model human judgment on a specific trait. In the past, the IEA has been used to accurately score a wide range of traits including the following:

- Overall quality
- Content
- Development
- Response to the prompt
- Effective sentences
- Focus and organization
- Grammar, usage, and mechanics (spelling)
- Word choice
- Development and details
- Conventions
- Focus
- Coherence
- Reading comprehension
- Progression of ideas
- Style
- Point of view
- Critical thinking
- Appropriate examples, reasons, and other evidence to support a position.
- Sentence structure
- Skilled use of language, and accurate and apt vocabulary

A number of these traits closely match those of the NAPLAN criteria, while other criteria are very closely related to features that are used within the analysis performed by IEA. As such, we applied our machine-learning-based approach, which adjusts its approach automatically to determine the features that best match each of the NAPLAN criteria. Because some of the features and methods for combining the features are proprietary, not all details on the specific features used in the models are described in this report.

	NAPLAN Online Writing Pilot Study Scoring Rubric	
	Marking Criterion	**Description of Marking Criterion**
1	Audience	The writer's capacity to orient, engage and persuade the reader
2	Text structure	The organization of the structural components of a persuasive text (introduction, body and conclusion) into an appropriate and effective text structure
3	Ideas	The selection, relevance and elaboration of ideas for a persuasive argument

4	Persuasive devices	The use of a range of persuasive devices to enhance the writer's position and persuade the reader
5	Vocabulary	The range and precision of contextually appropriate language choices
6	Cohesion	The control of multiple threads and relationships across the text, achieved through the use of grammatical elements (referring words, text connectives, conjunctions) and lexical elements (substitutions, repetitions, word associations)
7	Paragraphing	The segmenting of text into paragraphs that assists the reader to follow the line of argument
8	Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences
9	Punctuation	The use of correct and appropriate punctuation to aid the reading of the text
10	Spelling	The accuracy of spelling and the difficulty of the words used

Evaluating Responses for Scorability

Based on the essays on which it was trained, IEA can be set to have certain expectations about the content, style, quality, length, and skill level it expects to find in the writing it receives to score. If a new essay does not meet these expectations, then IEA can flag the essay for human review. IEA uses a variety of statistical and probabilistic checks to make this determination based on characteristics of the responses on which it was trained and experience with a variety of both good- and bad-faith responses. Responses may be flagged for the following reasons:

- The response may be too short to adequately evaluate the students' ability; in some circumstances, assigning the lowest possible score is the correct action to take in this case
- The response may be much longer than expected
- The response may be off-topic or it may be highly creative
- The response may not be in good faith (e.g., a refusal to write)
- The response may demonstrate a skill level that is very different from the expected skill level (this can happen when a 6^{th} grade student is asked to respond to a 12^{th} grade prompt and vice versa)
- The response may be in all capital letters and thus not demonstrate appropriate formal writing style.
- The response may include too much repeated content, such as copying and pasting the same paragraph over and over
- The response may not look like an essay. For example, it may be just a list of words or contain little to no punctuation.

IEA is able to assign a score to a majority of responses; the question is whether or not it should, and what the ramifications would be if it gives an inappropriate score. The thresholds that IEA uses to determine whether or not to report a score can be adjusted based on the nature of the assessment and the available testing data. For high stakes assessments, one generally wants to be more cautious. If the assessment is double-scored with one human score and one automated score, then one can flag fewer or next to no responses, relying on the human scorer, and potentially resolution, to identify any issues. In formative

settings where the goal is to provide immediate feedback and practice, it may be appropriate to score as many responses as possible, while at the same time flagging unusual responses for teacher review.

Across many types of operational assessments, IEA typically flags about two to five percent (2-5%) of responses using standard thresholds. We generally work with our customers to understand a given assessment scenario, discuss the tradeoffs and risks, and recommend the appropriate action for flagging unusual responses.

In the present analysis, we chose not to set any validation thresholds since the assumption was made that all essay input was appropriate. However, analyses could be performed which flag particular essays which may be suspect or which may not be scored as accurately by the automated model.

Evaluation of Scoring Engine Performance

The performance of a scoring model can be evaluated on how well the scores match human scoring, but also how well the scores align with the constructs of interest. The most common benchmark is to compute the reliability of the scoring engine by examining the agreement of IEA's predicted scores to human scorers, as compared to the agreement between human scorers. Metrics for computing the reliability include correlation, kappa, weighted kappa, and exact and adjacent agreement. Using "true scores" (e.g., the average of multiple scorers or the consensus score) for the comparison can provide more accurate measures of IEA's accuracy. In the present study, we provide measures of correlation, exact and adjacent agreement for the human-human reliability as well as for the IEA to the average of the human scorers.

Human agreement, however, is seldom sufficient as a means to evaluate performance. IEA performance can be compared against external variables that provide a measure of the validity of the scoring, including comparison of IEA scores with scores from concurrent administrations of tests with a similar construct, agreement with scores from subsequent tests, predicting student age or grade level, agreement to scorers with different levels of skill, and tests of scoring across different population subgroups. In the proposed project, we focused on reporting several agreement statistics: correlation and exact and adjacent agreement relative to the human marker performance.

Results from Analysis of the NAPLAN scripts

Analyses were conducted on 1356 scripts divided into samples 1, the training set which was used for calibrating the model, sample 2, a validation set, the predictions from which are used for the analysis presented here, and sample 3, a test set where human scores were not provided to KT and the predictions were sent directly to ACARA for their analysis. The counts in each set are as follows:

sample	1	2	3
count	677	340	339

Calibration of the models

Individual models were built for each of the ten criteria. For each criteria, the 677 scripts from sample 1 and the scores for the criteria were used to train the model. The model then contained the set of language features that best predict the human scores from sample 1. Because PKT uses proprietary language features in its modeling techniques, we do not provide the specific features used in each model. The features do fall into the categories of features described in the section and figures above.

Evaluation of performance on the validation set

The performance of a scoring model can be evaluated on how well the scores match human scoring, and also how well the scores align with the constructs of interest. The most common benchmark is to compute the reliability of the scoring engine by examining the agreement of IEA's predicted scores to human scorers, as compared to the agreement between human scorers. Metrics for computing the reliability include correlation, kappa, weighted kappa, and exact and adjacent agreement. Using "true scores" (e.g., the average of multiple scorers or the consensus score) for the comparison can provide more accurate measures of IEA's accuracy.

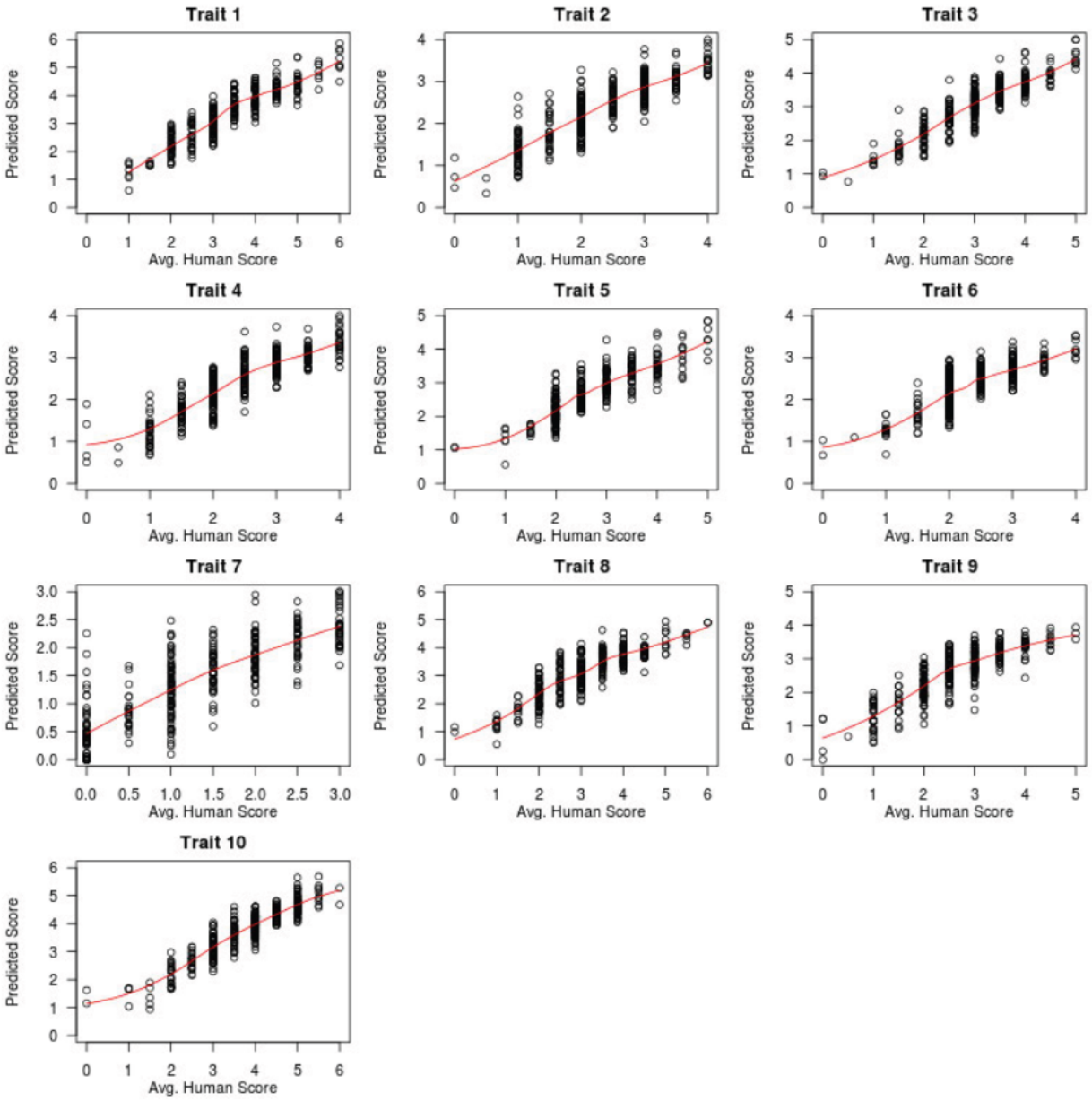
In the present analyses, we analyzed the performance on the 340 sample 2 scripts. We computed the correlation, exact and adjacent agreement for the two human raters as well as for the IEA's predicted score to the mean of the two human raters. The following table presents a summary of the performance of automated scoring. It presents agreement for H-H (human-human), IEA-H (IEA agreement to the average of the human scorers) and H1 or H2-IEA (IEA agreement to each of the individual human scorers).

Trait	H-H Cor	H-H Exact	H-H Adj	IEA-H Cor	IEA-H Exact	IEA-H Adj	h1 to IEA Cor	h1 to IEA Exact	h1 to IEA Adj	h2 to IEA Cor	h2 to IEA Exact	h2 to IEA Adj
1	0.82	61.2	97.6	0.90	61.8	99.7	0.87	62.1	99.1	0.85	60.6	98.8
2	0.81	68.8	99.7	0.86	65.9	99.1	0.83	67.6	98.8	0.81	62.4	99.7
3	0.78	62.4	99.1	0.89	65.9	99.4	0.83	64.1	99.7	0.85	68.2	99.4
4	0.73	55.6	98.5	0.87	64.1	99.4	0.80	61.8	99.4	0.82	63.2	99.7
5	0.76	64.4	97.6	0.86	65.9	99.7	0.80	62.1	98.5	0.81	65.6	98.8
6	0.71	71.8	99.7	0.80	73.5	100	0.74	72.1	100	0.74	70.0	100
7	0.83	69.7	99.4	0.82	57.6	98.5	0.78	55.6	98.5	0.79	60.0	98.5
8	0.78	58.5	95.9	0.85	60.6	97.9	0.81	59.7	98.5	0.80	57.1	96.2
9	0.76	60.6	97.9	0.84	60.0	99.1	0.78	55.0	98.5	0.79	63.2	98.2
10	0.84	66.2	98.8	0.92	67.6	99.4	0.88	65.6	99.1	0.88	67.9	99.4

Overall the results show that human-human agreement is generally high and in the range we typically see for careful scoring of scripts, with correlations running from .71 to .84 for different traits. The IEA correlation with the average of the human raters was also quite high, ranging from .80 to .92. Generally, the IEA to Human correlation was greater than that of the human-human correlation. This could be due

to the fact that the IEA is matching to the consensus (average of two raters). The results on the exact and adjacent agreement show similar patterns, with the IEA performance being very close to that of the human performance. We see slight decreases in IEA performance for criteria 7 (paragraph structure) for correlation and exact agreement, and for criteria 2 (text structure) for exact agreement. However, in each case, those differences are quite slight. These slight decrements in performance may be due to the fact that the IEA does not have many features that were developed explicitly for evaluating paragraph structure.

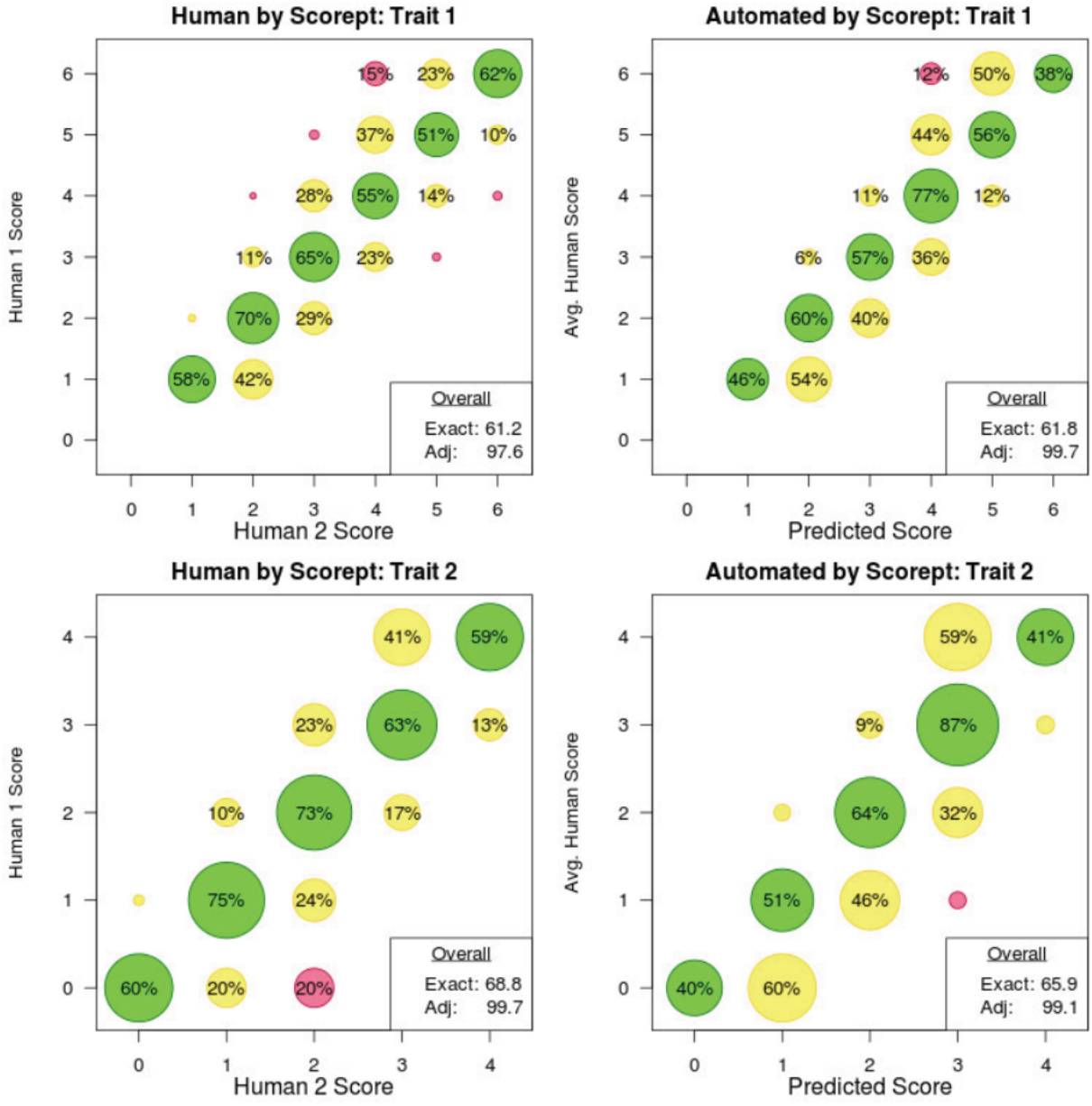
The IEA is able to provide a continuous score (e.g., decimal values) within the score range, although for computing agreement statistics and final scores, these continuous scores are rounded to integers. To provide a graphical indication of the agreement for each trait, we plotted the continuous predicted scores against the average human scores (which can be integers or fractions at 0.5s). These plots indicate that we see generally strong fits.

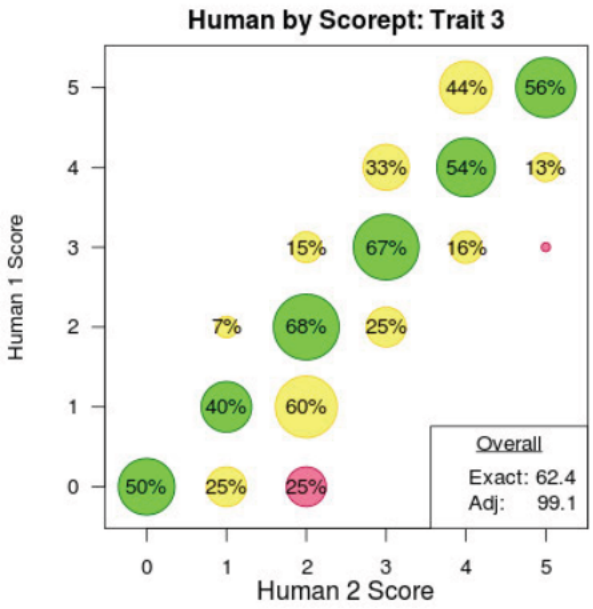


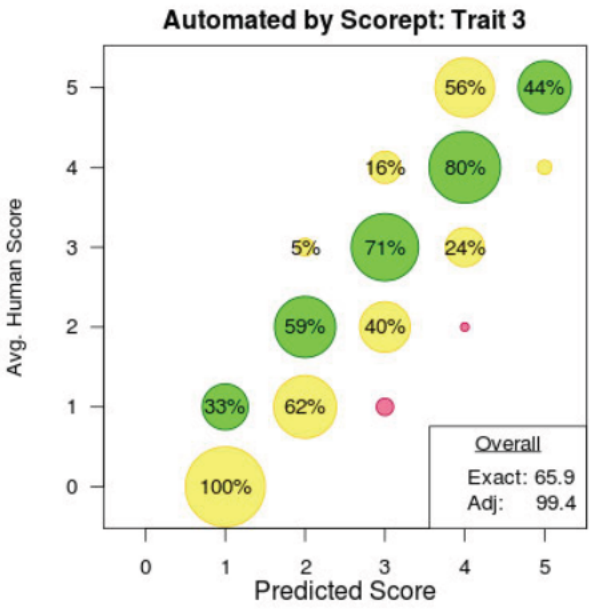
In order to evaluate how the distribution of scores provided by the human scorers compares to the automated scores, below we plot the agreement for each score point with separate plots for each of the ten criteria and for both human-human agreement and IEA-human agreement. The plots are followed by confusion matrices, which provide counts for each of the circles. Exact agreement is denoted in green, adjacent (within one score point) in yellow, and differences greater than one score point in red. These results indicate generally good agreement across the score points. We see some level of divergence, particularly at the lower score points and a bit at the higher score points. This effect is likely due to the fact that there were very few examples at the lower and upper ends (as can be seen in the confusion matrices that follow the plots), both in the sample 1 training set and in the sample 2 validation set.

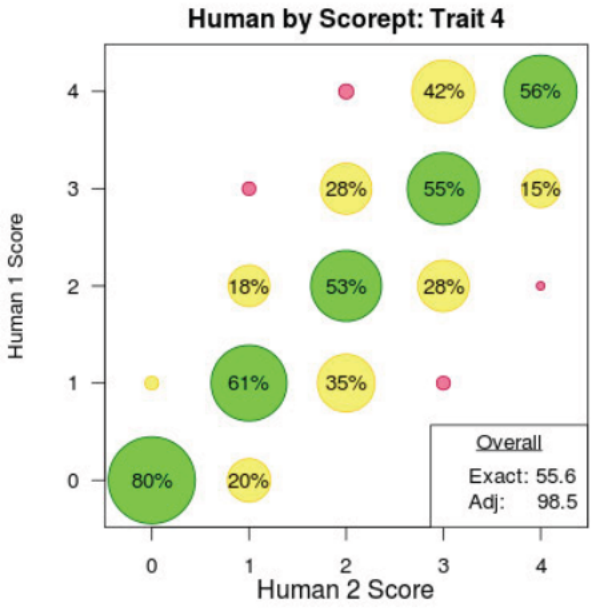
Without many samples in the training set, the system may not learn as well to generalize about the underrepresented score points, and thus you tend to see some level of regression towards the mean.

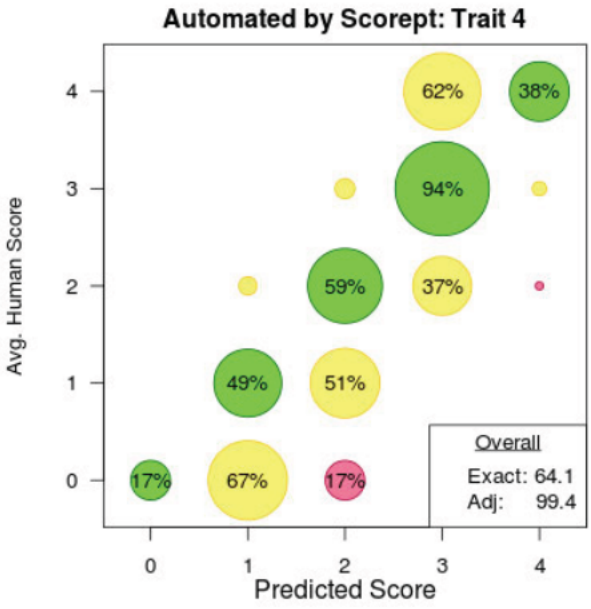
It is important to notice that although the percentages are shown, some of the percentages are based on very few score points. For example in trait 5, there were only 2 scripts at the lowest score point and only 7 that both scorers agreed were scored a 6. Thus, the tables following the graphs provide a different representation of the data, showing the raw score counts for each of the ten criteria.

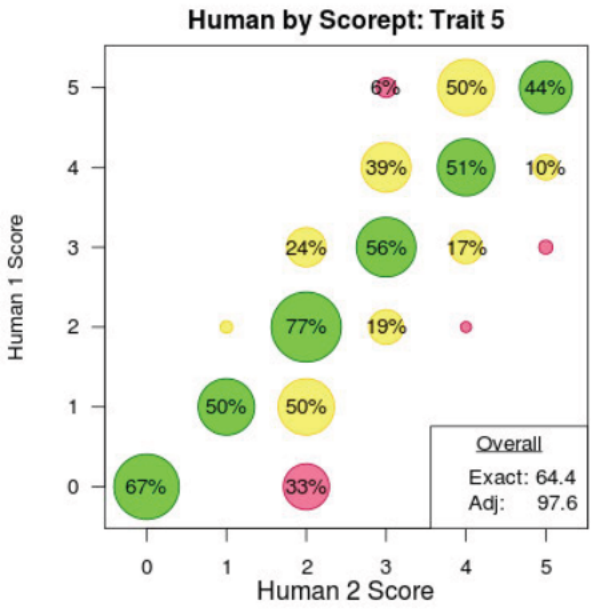


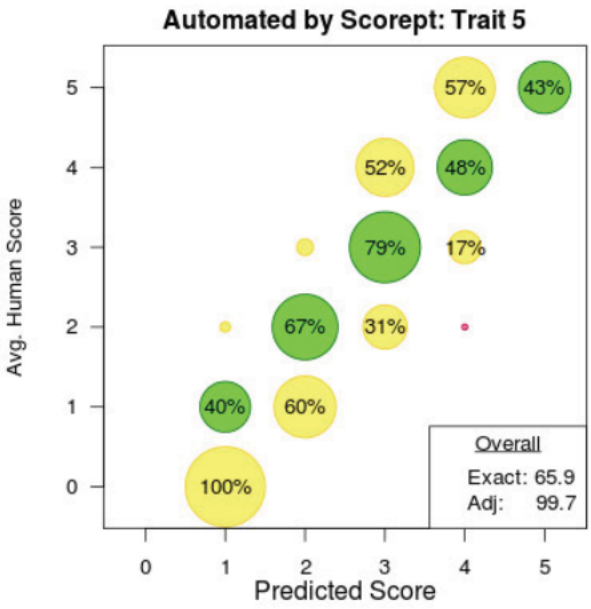

Human by Scorept: Trait 3
Human 1 Score
Human 2 Score
Overall
Exact: 62.4
Adj: 99.1

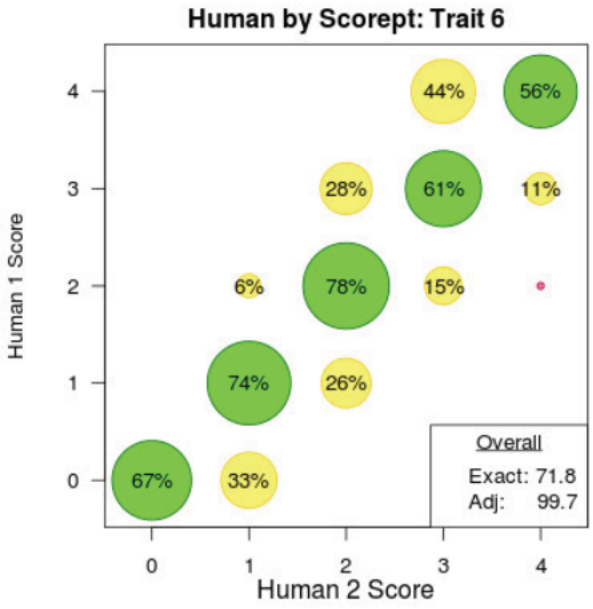

Automated by Scorept: Trait 3
Avg. Human Score
Predicted Score
Overall
Exact: 65.9
Adj: 99.4

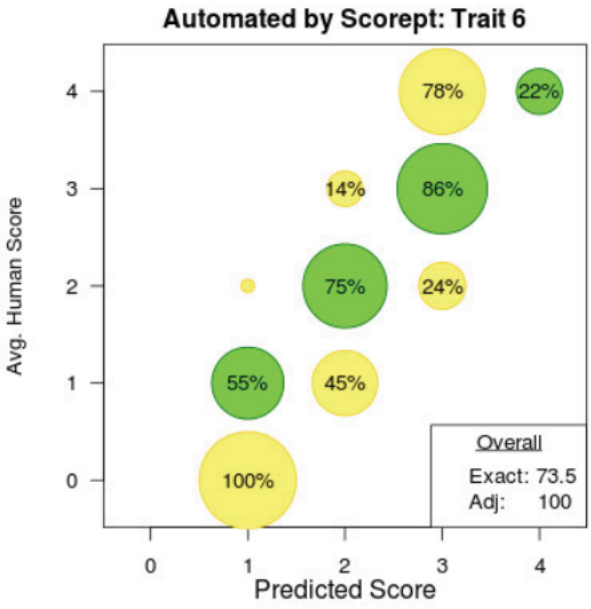

Human by Scorept: Trait 4
Human 1 Score
Human 2 Score
Overall
Exact: 55.6
Adj: 98.5

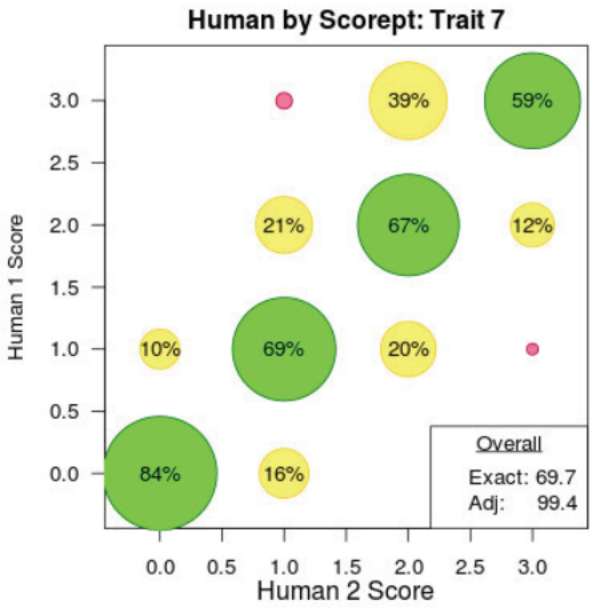

Automated by Scorept: Trait 4
Avg. Human Score
Predicted Score
Overall
Exact: 64.1
Adj: 99.4

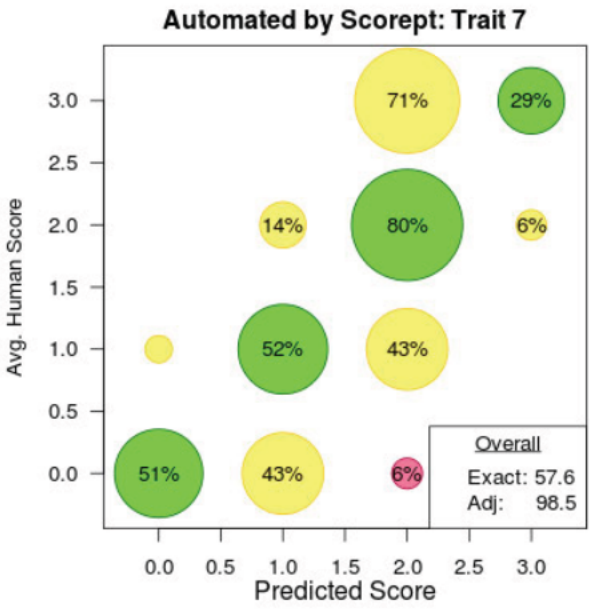

Human by Scorept: Trait 5
Human 1 Score
Human 2 Score
6%
50%
44%
39%
51%
10%
24%
56%
17%
77%
19%
50%
50%
67%
33%
Overall
Exact: 64.4
Adj: 97.6

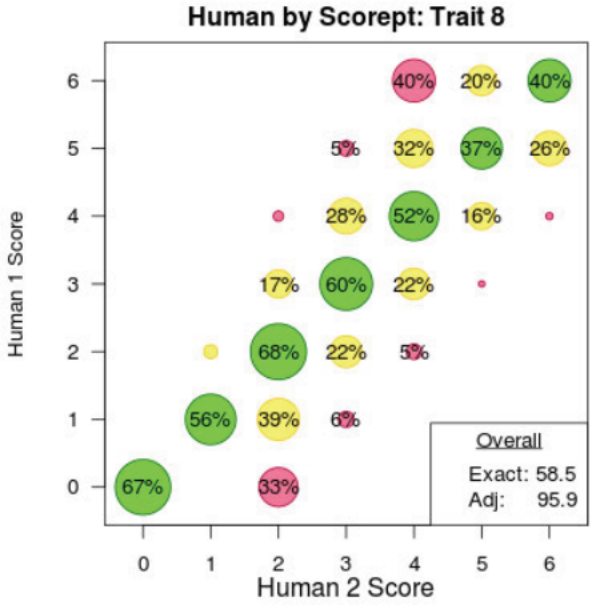

Automated by Scorept: Trait 5
Avg. Human Score
Predicted Score
57%
43%
52%
48%
79%
17%
67%
31%
40%
60%
100%
Overall
Exact: 65.9
Adj: 99.7

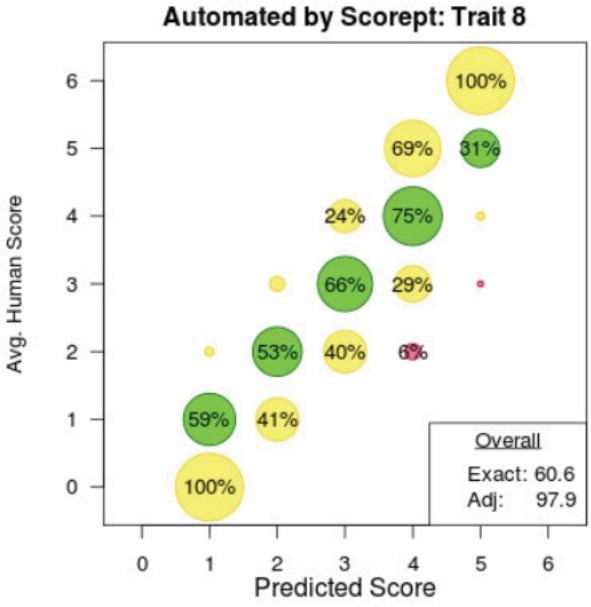

Human by Scorept: Trait 6
Human 1 Score
Human 2 Score
44%
56%
28%
61%
11%
6%
78%
15%
74%
26%
67%
33%
Overall
Exact: 71.8
Adj: 99.7

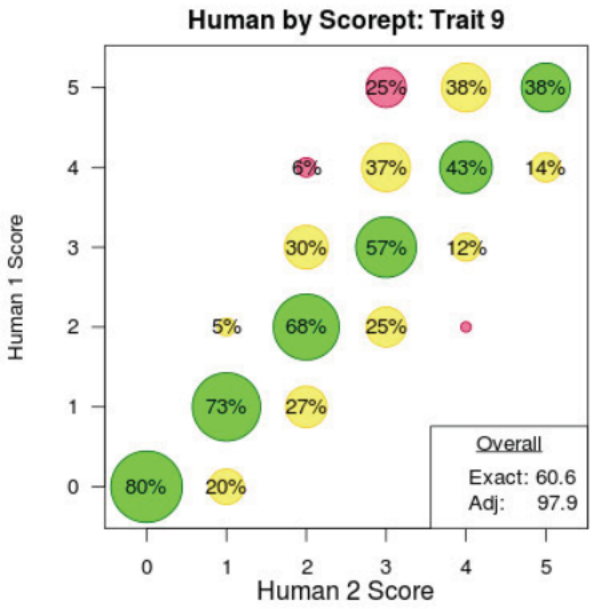

Automated by Scorept: Trait 6
Avg. Human Score
Predicted Score
78%
22%
14%
86%
75%
24%
55%
45%
100%
Overall
Exact: 73.5
Adj: 100

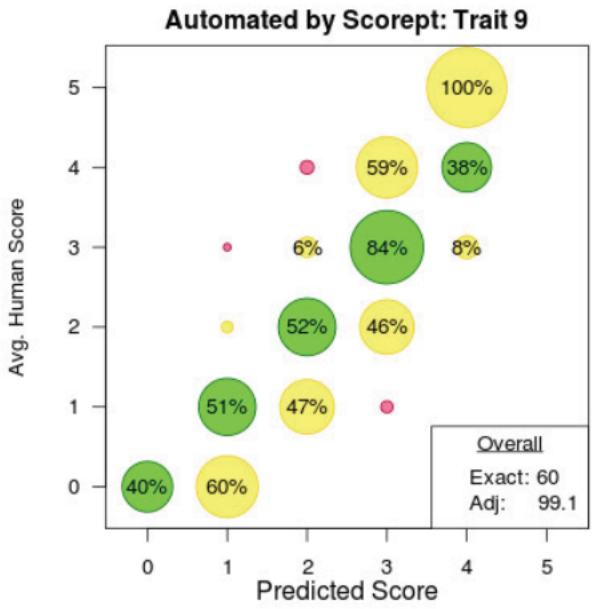

Human by Scorept: Trait 7
Human 1 Score
Human 2 Score
39%
59%
21%
67%
12%
10%
69%
20%
84%
16%
Overall
Exact: 69.7
Adj: 99.4

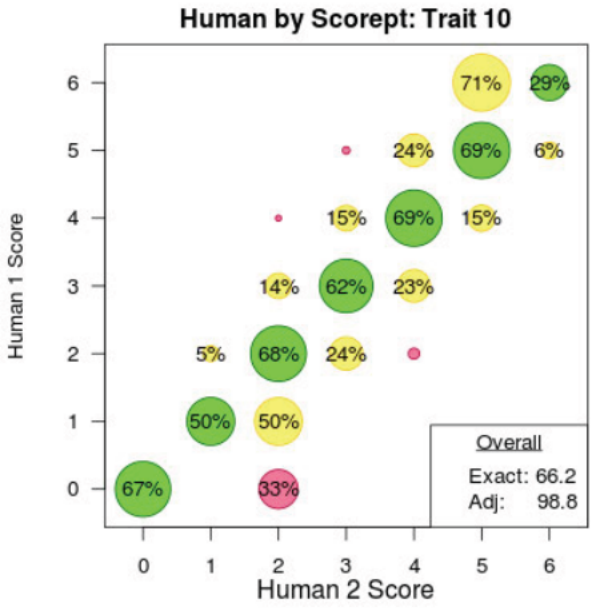

Automated by Scorept: Trait 7
Avg. Human Score
Predicted Score
71%
29%
14%
80%
6%
52%
43%
51%
43%
6%
Overall
Exact: 57.6
Adj: 98.5

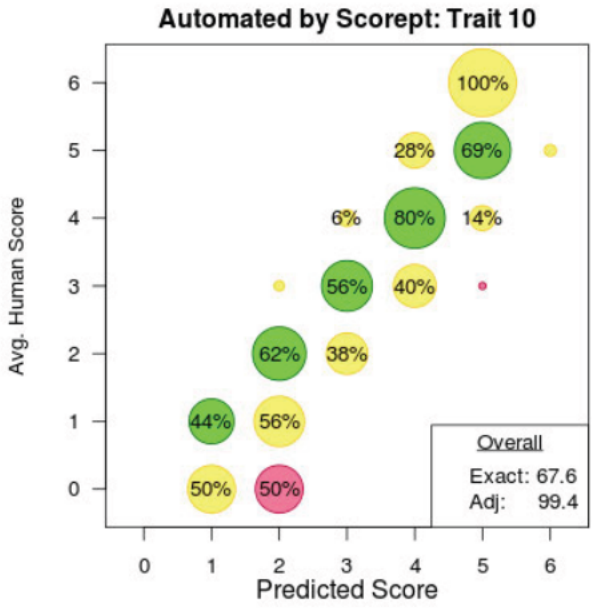

Human by Scorept: Trait 8
Human 1 Score
Human 2 Score
40%
20%
40%
5%
32%
37%
26%
28%
52%
16%
17%
60%
22%
68%
22%
5%
56%
39%
6%
67%
33%
Overall
Exact: 58.5
Adj: 95.9


Automated by Scorept: Trait 8
Avg. Human Score
Predicted Score
100%
69%
31%
24%
75%
66%
29%
53%
40%
6%
59%
41%
100%
Overall
Exact: 60.6
Adj: 97.9


Human by Scorept: Trait 9
Human 1 Score
Human 2 Score
25%
38%
38%
6%
37%
43%
14%
30%
57%
12%
5%
68%
25%
73%
27%
80%
20%
Overall
Exact: 60.6
Adj: 97.9


Automated by Scorept: Trait 9
Avg. Human Score
Predicted Score
100%
59%
38%
6%
84%
8%
52%
46%
51%
47%
40%
60%
Overall
Exact: 60
Adj: 99.1


Human by Scorept: Trait 10
Human 1 Score
Human 2 Score
71%
29%
24%
69%
6%
15%
69%
15%
14%
62%
23%
5%
68%
24%
50%
50%
67%
33%
Overall
Exact: 66.2
Adj: 98.8


Automated by Scorept: Trait 10
Avg. Human Score
Predicted Score
100%
28%
69%
6%
80%
14%
56%
40%
62%
38%
44%
56%
50%
50%
Overall
Exact: 67.6
Adj: 99.4

Trait 1

IEA

	Predicted							
Actual	0	1	2	3	4	5	6	total
0	0	0	0	0	0	0	0	0
1	0	6	7	0	0	0	0	13
2	0	0	48	32	0	0	0	80
3	0	0	8	71	45	0	0	124
4	0	0	0	9	64	10	0	83
5	0	0	0	0	14	18	0	32
6	0	0	0	0	1	4	3	8
total	0	6	63	112	124	32	3	340

Human

	h2							
h1	0	1	2	3	4	5	6	total
0	0	0	0	0	0	0	0	0
1	0	7	5	0	0	0	0	12
2	0	1	48	20	0	0	0	69
3	0	0	12	72	25	2	0	111
4	0	0	1	26	52	13	2	94
5	0	0	0	1	15	21	4	41
6	0	0	0	0	2	3	8	13
total	0	8	66	119	94	39	14	340

Trait 2

IEA

	Predicted					
Actual	0	1	2	3	4	total
0	2	3	0	0	0	5
1	0	41	37	3	0	81
2	0	5	89	45	0	139
3	0	0	9	85	4	98
4	0	0	0	10	7	17
total	2	49	135	143	11	340

Human

	H2					
H1	0	1	2	3	4	total
0	3	1	1	0	0	5
1	1	51	16	0	0	68
2	0	13	92	21	0	126
3	0	0	26	71	15	112
4	0	0	0	12	17	29
total	4	65	135	104	32	340

Trait 3

IEA

	Predicted						
Actual	0	1	2	3	4	5	total
0	0	3	0	0	0	0	3
1	0	7	13	1	0	0	21
2	0	0	51	35	1	0	87
3	0	0	8	113	38	0	159
4	0	0	0	10	49	2	61
5	0	0	0	0	5	4	9
total	0	10	72	159	93	6	340

Human

	H2						
H1	0	1	2	3	4	5	total
0	2	1	1	0	0	0	4
1	0	6	9	0	0	0	15
2	0	5	46	17	0	0	68
3	0	0	24	107	26	2	159
4	0	0	0	26	42	10	78
5	0	0	0	0	7	9	16
total	2	12	80	150	75	21	340

Trait 4

IEA

	Predicted					
Actual	0	1	2	3	4	total
0	1	4	1	0	0	6
1	0	36	38	0	0	74
2	0	5	84	52	1	142
3	0	0	4	88	2	94
4	0	0	0	15	9	24
total	1	45	127	155	12	340

Human

	H2					
H1	0	1	2	3	4	total
0	4	1	0	0	0	5
1	1	31	18	1	0	51
2	0	25	72	38	1	136
3	0	2	29	58	16	105
4	0	0	1	18	24	43
total	5	59	120	115	41	340

Trait 5

IEA

	Predicted						
Actual	0	1	2	3	4	5	total
0	0	2	0	0	0	0	2
1	0	6	9	0	0	0	15
2	0	3	121	55	1	0	180
3	0	0	4	74	16	0	94
4	0	0	0	22	20	0	42
5	0	0	0	0	4	3	7
total	0	11	134	151	41	3	340

Human

	H2						
H1	0	1	2	3	4	5	total
0	2	0	1	0	0	0	3
1	0	5	5	0	0	0	10
2	0	4	125	31	3	0	163
3	0	0	24	55	17	3	99
4	0	0	0	19	25	5	49
5	0	0	0	1	8	7	16
total	2	9	155	106	53	15	340

Trait 6

IEA

	Predicted					
Actual	0	1	2	3	4	total
0	0	3	0	0	0	3
1	0	17	14	0	0	31
2	0	4	161	51	0	216
3	0	0	11	70	0	81
4	0	0	0	7	2	9
total	0	24	186	128	2	340

Human

	H2					
H1	0	1	2	3	4	total
0	2	1	0	0	0	3
1	0	14	5	0	0	19
2	0	12	157	30	1	200
3	0	0	29	62	11	102
4	0	0	0	7	9	16
total	2	27	191	99	21	340

Trait 7

IEA

	Predicted				
Actual	0	1	2	3	total
0	41	35	5	0	81
1	6	64	53	0	123
2	0	14	81	6	101
3	0	0	25	10	35
total	47	113	164	16	340

Human

	H2				
H1	0	1	2	3	total
0	58	11	0	0	69
1	12	80	23	1	116
2	0	20	64	12	96
3	0	1	23	35	59
total	70	112	110	48	340

Trait 8

IEA

	Predicted							
Actual	0	1	2	3	4	5	6	total
0	0	2	0	0	0	0	0	2
1	0	13	9	0	0	0	0	22
2	0	2	56	42	6	0	0	106
3	0	0	6	82	36	1	0	125
4	0	0	0	16	50	1	0	67
5	0	0	0	0	11	5	0	16
6	0	0	0	0	0	2	0	2
total	0	17	71	140	103	9	0	340

Human

	H2							
H1	0	1	2	3	4	5	6	total
0	2	0	1	0	0	0	0	3
1	0	10	7	1	0	0	0	18
2	0	4	64	21	5	0	0	94
3	0	0	20	69	25	1	0	115
4	0	0	2	24	45	14	1	86
5	0	0	0	1	6	7	5	19
6	0	0	0	0	2	1	2	5
total	2	14	94	116	83	23	8	340

Trait 9

IEA

	Predicted						
Actual	0	1	2	3	4	5	total
0	2	3	0	0	0	0	5
1	0	22	20	1	0	0	43
2	0	3	76	68	0	0	147
3	0	1	7	91	9	0	108
4	0	0	1	20	13	0	34
5	0	0	0	0	3	0	3
total	2	29	104	180	25	0	340

Human

	H2						
H1	0	1	2	3	4	5	total
0	4	1	0	0	0	0	5
1	0	27	10	0	0	0	37
2	0	6	81	30	2	0	119
3	0	0	36	69	15	0	120
4	0	0	3	19	22	7	51
5	0	0	0	2	3	3	8
total	4	34	130	120	42	10	340

Trait 10

IEA

	Predicted							
Actual	0	1	2	3	4	5	6	total
0	0	1	1	0	0	0	0	2
1	0	4	5	0	0	0	0	9
2	0	0	28	17	0	0	0	45
3	0	0	2	49	35	1	0	87
4	0	0	0	8	104	18	0	130
5	0	0	0	0	18	45	2	65
6	0	0	0	0	0	2	0	2
total	0	5	36	74	157	66	2	340

Human

	H2							
H1	0	1	2	3	4	5	6	total
0	2	0	1	0	0	0	0	3
1	0	3	3	0	0	0	0	6
2	0	2	25	9	1	0	0	37
3	0	0	11	48	18	0	0	77
4	0	0	1	19	90	20	0	130
5	0	0	0	1	19	55	5	80
6	0	0	0	0	0	5	2	7
total	2	5	41	77	128	80	7	340

Results on Sample 3

The scripts from Sample 3 were scored blindly by the IEA and the results were returned to ACARA. Near the end of this project, the human marker scores were provided for the 339 scripts. Below is a summary of the performance of the human markers and the IEA as measured by correlation, exact and adjacent agreement. The results show that IEA is able to provide scores that are at or above the level of the human-human reliability rates

Trait	H-H Cor	H-H Exact	H-H Adj	IEA-H Cor	IEA-H Exact	IEA-H Adj	h1 to IEA Cor	h1 to IEA Exact	h1 to IEA Adj	h2 to IEA Cor	h2 to IEA Exact	h2 to IEA Adj
1	0.83	64.3	97.6	0.90	65.2	99.1	0.86	58.7	98.2	0.87	62.5	99.4
2	0.81	68.4	99.7	0.89	73.5	100.0	0.84	69.6	99.4	0.85	70.2	100.0
3	0.79	65.5	98.5	0.87	64.3	98.8	0.82	63.7	98.8	0.83	65.5	98.8
4	0.74	60.2	97.9	0.88	65.5	99.7	0.82	64.3	99.7	0.81	65.8	99.4
5	0.76	64.3	97.6	0.87	69.6	99.4	0.82	68.1	99.4	0.81	69.3	99.1
6	0.67	69.0	98.8	0.77	70.2	99.7	0.72	68.7	99.4	0.70	68.1	99.7
7	0.79	64.3	97.9	0.88	63.4	98.8	0.85	63.4	99.1	0.81	58.1	99.1
8	0.79	59.3	97.9	0.84	60.5	97.9	0.79	58.1	97.3	0.80	59.6	97.9
9	0.76	64.0	97.6	0.84	60.2	99.4	0.80	60.2	98.8	0.77	59.3	99.1
10	0.85	68.7	97.9	0.88	65.5	98.5	0.86	64.6	98.2	0.84	63.4	99.1

Conclusions

Overall, the results from the analyses are encouraging. All ten of the NAPLAN scoring criteria had levels of human-human reliability that made them amenable to automated scoring. IEA was able to closely match, and in some times exceed the agreement rates of the human scorers. The performance was fairly comparable across the 10 different scoring criteria, with slight decrements in performance for the criteria that use structural or paragraphing elements. The results indicated that the model built on Sample 1 showed strong generalization to marking scripts from Samples 2 and 3.

IEA's approach to predicting scores is to intuit a set of construct-relevant features and how to weight them based on modeling the human scores. However, not all features in the description of the criteria are fully reflected in the features used by IEA. For example, IEA doesn't take into account some aspects such as measuring appropriateness of paragraph breaks. However, it does have features that reflect organizational structure and coherence that are related to the same criteria. Ongoing work can be performed to refine some of these features so that they more closely match the criteria that are used by the human scorers. Additional work can also be performed to improve the scoring accuracy. For example, additional training data could be used in order to improve scoring accuracy at the ends of the scale. For instance, for the scoring of the 339 sample three scripts, the scripts from both samples one and two could be used for training. This would produce greater numbers of examples of scripts at both the high end low ends of the scoring scales. Finally, additional work could be performed to change the thresholds of the IEA generated continuous scores. Adjusting thresholds allows adjusting the scoring distribution to better match the scoring distribution of the human scores and could likely improve the exact and adjacent agreements slightly beyond their current level of performance.

References

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John. ed. "Latent Dirichlet allocation". *Journal of Machine Learning Research* 3

Bernstein, J. Taussig, K and Godfrey, J (1994) *MACROPHONE.* Linguistic Data Consortium, Philadelphia.

Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407

Foltz, P. W. & Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, *35*(12), 51-60.

Foltz, P. W., Streeter, L. E., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. In M. Shermis and J. Burstein, (Eds.). *The Handbook of Automated Essay Evaluation: Current Applications and New Directions.* New York: Routledge.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211-244

Hearst, M., Kukich, K., Hirschman, L., Breck, E., Light, M., Burge,J., Ferro, L., Landauer, T. K., Laham, D., and Foltz, P. W., The Debate on Automated Essay Grading, in IEEE Intelligent Systems (Sept/Oct 2000).

Hill, W., Stead, L., Furnas, G., and Rosenstein, M. Recommending and Evaluating Choices in a Virtual Community of Use. CHI'95. May 7-11,

Landauer, T.K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, 15(1), 92-108.

Streeter, L. A. and Lochbaum, K. E. (1988): "An Expert/Expert Locating System based on Automatic Representation of Semantic Structure". In *Proceedings of the Fourth IEEE Conference on Artificial Intelligence Applications,* Computer Society of the IEEE, San Diego, CA, pp. 345 – 349.

Williamson, D. M., Bennett, R., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D., Way, D., and Sweeney, K. (2010, June). *Automated Scoring for the Assessment of Common Core Standards.*

The Performance of the CRASE™ Engine on the NAPLAN Writing Prompt Scoring

Pacific Metrics Corporation
1 Lower Ragsdale Drive
Building 1, Suite 150
Monterey, CA 93940

Table of Contents

INTRODUCTION AND OVERVIEW

This report outlines the steps and results of the proof-of-concept automated essay scoring work conducted for the Australian Curriculum, Assessment and Reporting Authority (ACARA) for the *National Assessment Program – Literacy and Numeracy (NAPLAN) Online Trial Study 2013: Automated Essay Scoring of Writing Scripts and Report.* The purpose of this study was to evaluate the accuracy of Pacific Metric's automated scoring engine, CRASE™, on all ten marking criteria applied to the scoring of a single writing prompt. The document begins with a brief description of the writing guide, and of the CRASE scoring engine and its alignment to the writing guide. The process used for training the engine is described and the results of that training applied to a cross-validation sample are illustrated in a series of tables and evaluated using various statistics. Finally, the results of applying the scoring model to the blind evaluation sample are shown.

BACKGROUND

NAPLAN ONLINE TRIAL STUDY

The NAPLAN program administered the same persuasive writing prompt to students in Years 3, 5, 7, and 9. The test was administered online. As a result, the student responses reflect what the student typed into the platform, including HTML-based formatting made available in the test delivery client. The student responses were scored using ten marking criteria in the Persuasive Writing Marking Guide. For the purposes of this study, the student responses were independently scored by two human raters blind to the grade of the student.

A description of the Persuasive Writing Marking Guide used in scoring appears in Table 1. More detail on the score point descriptions for each of the 10 marking criteria can be found in the marking guide (Accessible at: http://www.nap.edu.au/verve/resources/Amended_2013_Persuasive_Writing_Marking_Guide_-With_cover.pdf).

Table 1. Description of Ten Marking Criteria and Score Ranges in the NAPLAN Online Study

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

Note. Table was adapted from page 6 in the NAPLAN Persuasive Writing Marking Guide (Accessible at: http://www.nap.edu.au/verve/_resources/Amended_2013_Persuasive_Writing_Marking_Guide_-With_cover.pdf).

CRASE DESCRIPTION

At its most basic level, scoring is an exercise in categorization. Using experience, along with training materials, marking rubrics, etc., a human rater classifies a student's response into one of several defined categories or scores. CRASE analyzes a sample of already-scored student response to produce a model of the human raters' scoring behavior. In general, the system will score as reliably as the sample from which the scoring models are built. By emulating human scoring behavior, CRASE essentially predicts the score (or scores in the case of multiple-dimension marking rubrics) that a human rater would assign to a given student response.

In training the engine and scoring responses, CRASE utilizes a sequential process to first analyze and then score students' responses. When a response is submitted to the engine, it moves through three phases in the scoring process: identifying non-attempts, feature extraction, and scoring.

- ***Identifying Non-Attempt and '0' Scores***. The response is first reviewed by the system to determine whether it is a valid attempt at the item and/or whether the response will earn a score of '0' as assigned in the writing guide. If it is not a valid attempt (e.g., it is blank or gibberish) or does not satisfy minimal criteria to earn a score of '1' or greater, the script is flagged and removed from the remaining feature extraction and scoring process.
- ***Extraction of Features***. If it is a valid attempt, the response is submitted to one of the feature extraction engines. In this phase, a vector of values is generated that represents both the marking guide and the construct the item is intended to assess.
- ***Predicting a Score***. The vector of values is submitted to a scoring engine that uses a statistical model and/or a series of computational linguistic procedures to classify the script into a score category. It is at this stage that the model derived from the rater sample is applied to predict the score a rater would provide. The predicted score and any non-attempt flags are then returned.

The feature extraction stage begins with preprocessing student responses by tokenizing elements in the response, counting basic text elements (e.g., number of words, sentences, paragraphs), and producing various representations of the response (e.g., part of speech tagging, spell-correction). The preprocessed responses are then submitted to the feature extraction engines. These functions are developed to represent key writing characteristics of essay responses. (Examples of functions include identification of usage and mechanics errors typically seen in student essays, a measure of variation in sentence type, extent of tone and personal engagement in phrasing, and the use of developmental phrasing.) These functions are applied to the processed response to produce one or more variables that represent various writing features.

Following feature extraction, the numerical feature values are entered in a statistical scoring model (e.g., regression, gradient boosted machines) for scoring. During the engine training

phase, the parameters used to predict scores are estimated using the training sample. After evaluation of the results on the training sample, the parameters are adjusted as needed. The final set of parameters estimated during training is used to predict marking criterion scores.

Table 2 presents a proposed alignment between the CRASE features extracted from the essays and the marking criteria. The scoring model considers all features in the prediction of the score for any criterion. The weighting of the features will vary for each marking criterion. Thus, a feature aligned to 'Audience' may also be used in the model to predict a score in 'Ideas' and generally will be weighted differently. This feature overlap in scoring generally occurs in multi-trait writing scoring because the traits tend to be highly correlated with one another.

Table 2. Description of Ten Marking Criteria and the Alignment of the CRASE Feature Extraction

Marking Criterion	Description of Marking Criterion	CRASE Feature
Audience	The writer's capacity to orient, engage and persuade the reader	Words that evoke tone, mood, personal engagement; Sentiment analysis; Active/passive voice; Informal language usage
Text structure	The organisation of the structural components of a persuasive text (introduction, body and conclusion) into an appropriate and effective text structure	Discourse phrasing and transition words
Ideas	The selection, relevance and elaboration of ideas for a persuasive argument	Bag of words methods that use term document frequency matrix; Topic identification
Persuasive devices	The use of a range of persuasive devices to enhance the writer's position and persuade the reader	Words that evoke tone, mood, personal engagement in topic; Bag of words methods
Vocabulary	The range and precision of contextually appropriate language choices	Overly-common, unique, and informal words; Part of speech usage
Cohesion	The control of multiple threads and relationships across the text, achieved through the use of grammatical elements (referring words, text connectives, conjunctions) and lexical elements (substitutions, repetitions, word associations)	Bag of words methods at sentence and paragraph level
Paragraphing	The segmenting of text into paragraphs that assists the reader to follow the line of argument	Number of paragraphs; Paragraph length; Bag of words methods at sentence and paragraph level
Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences	Sentence beginnings, lengths, and type; Punctuation use; Common usage and style errors; Style errors
Punctuation	The use of correct and appropriate punctuation to aid the reading of the text	Existence of punctuation; Capitalization use; Punctuation error detection; Named entity recognition
Spelling	The accuracy of spelling and the difficulty of the words used	Standard spell-correction methods

METHODS

DATA HANDLING AND CLEANING

Pacific Metrics received an Excel spreadsheet with records containing an anonymous student identifier, a sample designator (1, 2, 3), scores in each marking criterion for two human rater scores for samples 1 and 2 only, and the total score (for samples 1 and 2 only -- summed across the ten marking criteria scores for each rater). The file contained 1356 records. In addition, 1525 HTML files, each containing a single student response with the filename keyed to the student identifier, were provided.

The individual response files were merged into a single text data file that was keyed by the student identifier. Student responses were processed by converting </p>, </P>, , and header tags to return characters, replacing special characters (for example, &) to their single character counterparts, and then stripping the remaining HTML tags from the response. The aggregated file was then merged with the Excel file using the student identifier. One hundred sixty-nine of the student text responses were not merged with the scores because there was no matching ID between the HTML-response files and the provided Excel file. All IDs in the original Excel file had a student response or script associated with them. As a result, a total of 1356 records were available for engine training, cross-validation, and blind scoring.

In addition, Pacific Metrics verified that the scores provided by the humans were within the valid range provided in the marking guide for the first and second human raters. All scores were in the acceptable range with the exception of one score supplied by the first human rater in the "Ideas" marking criterion, which was a score of '9'.

DATA REVIEW

Pacific Metrics used the sample designations provided in the data file for training and cross-validation after ensuring that the mean scores across the two samples for each marking criteria were not statistically or practically significantly different under the assumption that the samples were randomly generated by ACARA staff from the same population of students. Table 3 presents the student counts for each of the samples. Pacific Metrics used the responses designated as sample 1 as the "Training Sample," and this set was comprised of 677 responses. Pacific Metrics used the responses designated as sample 2 as a "Cross-Validation Sample," and this set was comprised of 340 responses. For each of sample 1 and sample 2, there were two human-assigned scores in each of the ten marking criterion. The sample designation 3 included 339 responses for which no human scores were provided. In this report, the term "Blind Evaluation Sample" is used to describe this sample.

Table 3. Samples Used in Engine Training and Evaluation

Sample Designation	Sample Description	N	Percent
1	Training Sample	677	50%
2	Cross-Validation Sample	340	25%
3	Blind Evaluation Sample	339	25%

Tables 4a through 4c present the score frequency distributions, the means and standard deviations, and agreement indices (exact, adjacent, non-adjacent, Kappa, Quadratic Weighted Kappa [QWK], and correlation) for each marking criteria for each of the two human rater scores for the training sample. A review of the score point distributions shows that the full range of valid scores in the marking guides appeared for both human rater scores, although some scores had very few responses (namely at the ends of the score ranges). The exact agreement rates between the two human raters' scores ranged across the marking criteria from a low of 58.3% (Sentence Structure) to a high of 70.0% (Cohesion). The non-adjacent rates were quite low ranging from 0.4% (Text Structure) to 3.8% (Punctuation). The Kappa values ranged from .430 (Persuasive Devices) to .572 (Spelling). The Quadratic Weighted Kappa values ranged from .674 (Cohesion) to .854 (Spelling).

Table 4a. **Training Sample** -- Score Point Distributions (SPDs), Means and Standard Deviations, and Agreement Indices for Each Human Rater (H1, H2) for First Four Marking Criteria

	Audience		Text Structure		Ideas		Persuasive Devices	
Score	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**
0	0.4%	0.4%	0.7%	0.9%	0.7%	0.7%	0.9%	1.3%
1	3.0%	2.8%	19.8%	18.9%	4.6%	3.6%	17.7%	15.5%
2	20.1%	18.0%	38.1%	37.1%	20.4%	20.4%	40.8%	35.3%
3	35.9%	36.3%	33.4%	31.6%	49.9%	46.8%	31.5%	35.9%
4	28.2%	26.0%	8.0%	11.5%	18.8%	22.3%	9.2%	12.0%
5	8.7%	11.7%	.	.	5.5%	6.2%	.	.
6	3.7%	4.7%	.	.	.	.	.	.
Mean	3.29	3.39	2.28	2.34	2.98	3.05	2.30	2.42
SD	1.11	1.16	0.90	0.94	0.93	0.94	0.90	0.93
Agreement	**H1-H2**		**H1-H2**		**H1-H2**		**H1-H2**	
Exact	62.6%		65.9%		68.3%		59.8%	
Adj.	35.2%		33.7%		31.1%		38.8%	
Non-A.	2.2%		0.4%		0.6%		1.3%	
Kappa	.500		.517		.533		.430	
QWK	.829		.790		.809		.738	
Correlation	.833		.793		.811		.745	

Note. One training sample response received a score of '9' in 'Ideas. This response was removed from the sample for the calculations appearing in this table.

Table 4b. **Training Sample** -- Score Point Distributions, Means and Standard Deviations, and Agreement Indices for Each Human Rater (H1, H2) for Second Four Marking Criteria

	Vocabulary		Cohesion		Paragraphing		Sentence Structure	
Score	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**
0	0.6%	0.7%	0.7%	0.7%	21.1%	19.9%	0.9%	0.9%
1	3.4%	2.4%	7.4%	8.3%	31.3%	31.6%	5.9%	4.3%
2	49.3%	46.2%	59.8%	56.1%	32.1%	31.5%	24.5%	26.3%
3	31.0%	29.3%	27.8%	28.1%	15.5%	17.0%	36.8%	34.9%
4	12.1%	15.6%	4.3%	6.8%	.	.	24.4%	24.5%
5	3.6%	5.5%	.	.	.	.	6.7%	6.9%
6	.	.	.	.	.	.	0.9%	2.2%
Mean	2.61	2.74	2.27	2.32	1.42	1.45	3.01	3.08
SD	0.89	0.97	0.69	0.75	0.99	0.99	1.07	1.11
Agreement	**H1-H2**		**H1-H2**		**H1-H2**		**H1-H2**	
Exact	64.8%		70.0%		65.9%		58.3%	
Adj.	33.4%		28.7%		33.4%		38.3%	
Non-A.	1.8%		1.3%		0.7%		3.4%	
Kappa	.467		.481		.534		.437	
QWK	.768		.674		.815		.780	
Correlation	.777		.677		.816		.781	

Table 4c. **Training Sample** -- Score Point Distributions, Means and Standard Deviations, and Agreement Indices for Each Human Rater (H1, H2) for Last Two Marking Criteria

	Punctuation		Spelling	
Score	**H1**	**H2**	**H1**	**H2**
0	1.6%	2.1%	0.4%	0.6%
1	12.3%	10.9%	2.5%	1.9%
2	34.6%	31.8%	13.9%	11.7%
3	34.3%	35.0%	25.0%	26.0%
4	15.8%	17.7%	32.6%	32.5%
5	1.5%	2.5%	23.2%	25.6%
6	.	.	2.4%	1.8%
Mean	2.55	2.63	3.66	3.72
SD	1.00	1.04	1.15	1.11
Agreement	**H1-H2**		**H1-H2**	
Exact	60.6%		67.8%	
Adj.	35.6%		30.7%	
Non-A.	3.8%		1.5%	
Kappa	.458		.572	
QWK	.755		.854	
Correlation	.758		.855	

Tables 5a through 5c present the score frequency distributions, the means and standard deviations, and agreement indices (exact, adjacent, non-adjacent, Kappa, QWK, and correlation) for each marking criteria for each of the two human rater scores for the cross-validation sample.

As with the training sample, the full range of valid scores in the marking guides appeared for both human rater scores, and some scores had very few responses at the ends of the score ranges. The exact agreement rates between the two human rater scores ranged across the marking criteria from a low of 55.6% (Persuasive Devices) to a high of 71.8% (Cohesion). The non-adjacent rates were quite low ranging from 0.3% (Text Structure and Cohesion) to 4.1% (Sentence Structure). The Kappa values ranged from .377 (Persuasive Devices) to .585 (Paragraphing). The Quadratic Weighted Kappa values ranged from .713 (Cohesion) to .844 (Spelling).

Table 5a. **Cross-Validation Sample --** Score Point Distributions, Means and Standard Deviations, and Agreement Indices for Each Human Rater (H1, H2) for First Four Marking Criteria

	Audience		**Text Structure**		**Ideas**		**Persuasive Devices**	
Score	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**
0	0.0%	0.0%	1.5%	1.2%	1.2%	0.6%	1.5%	1.5%
1	3.5%	2.4%	20.0%	19.1%	4.4%	3.5%	15.0%	17.4%
2	20.3%	19.4%	37.1%	39.7%	20.0%	23.5%	40.0%	35.3%
3	32.7%	35.0%	32.9%	30.6%	46.8%	44.1%	30.9%	33.8%
4	27.7%	27.7%	8.5%	9.4%	22.9%	22.1%	12.7%	12.1%
5	12.1%	11.5%	.	.	4.7%	6.2%	.	.
6	3.8%	4.1%	.	.	.	.	.	.
Mean	3.36	3.39	2.27	2.28	3.00	3.02	2.38	2.38
SD	1.15	1.12	0.93	0.92	0.95	0.95	0.94	0.96
Agreement	**H1-H2**		**H1-H2**		**H1-H2**		**H1-H2**	
Exact	61.2%		68.8%		62.4%		55.6%	
Adj.	36.5%		30.9%		36.8%		42.9%	
Non-A.	2.4%		0.3%		0.9%		1.5%	
Kappa	.485		.558		.456		.377	
QWK	.821		.811		.776		.727	
Correlation	.822		.811		.776		.727	

Table 5b. **Cross-Validation Sample --** Score Point Distributions, Means and Standard Deviations, and Agreement Indices for Each Human Rater (H1, H2) for Second Four Marking Criteria

	Vocabulary		Cohesion		Paragraphing		Sentence Structure	
Score	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**	**H1**	**H2**
0	0.9%	0.6%	0.9%	0.6%	20.3%	20.6%	0.9%	0.6%
1	2.9%	2.7%	5.6%	7.9%	34.1%	32.9%	5.3%	4.1%
2	47.9%	45.6%	58.8%	56.2%	28.2%	32.4%	27.7%	27.7%
3	29.1%	31.2%	30.0%	29.1%	17.4%	14.1%	33.8%	34.1%
4	14.4%	15.6%	4.7%	6.2%	.	.	25.3%	24.4%
5	4.7%	4.4%	.	.	.	.	5.6%	6.8%
6	.	.	.	.	.	.	1.5%	2.4%
Mean	2.67	2.72	2.32	2.32	1.43	1.40	3.00	3.07
SD	0.95	0.93	0.69	0.73	1.00	0.97	1.09	1.10
Agreement	**H1-H2**		**H1-H2**		**H1-H2**		**H1-H2**	
Exact	64.4%		71.8%		69.7%		58.5%	
Adj.	33.2%		27.9%		29.7%		37.4%	
Non-A.	2.4%		0.3%		0.6%		4.1%	
Kappa	.465		.509		.585		.440	
QWK	.759		.713		.834		.776	
Correlation	.760		.714		.835		.777	

Table 5c. **Cross-Validation Sample --** Score Point Distributions, Means and Standard Deviations, and Agreement Indices for Each Human Rater (H1, H2) for Last Two Marking Criteria

	Punctuation		Spelling	
Score	**H1**	**H2**	**H1**	**H2**
0	1.5%	1.2%	0.9%	0.6%
1	10.9%	10.0%	1.8%	1.5%
2	35.0%	38.2%	10.9%	12.1%
3	35.3%	35.3%	22.7%	22.7%
4	15.0%	12.4%	38.2%	37.7%
5	2.4%	2.9%	23.5%	23.5%
6			2.1%	2.1%
Mean	2.59	2.56	3.74	3.74
SD	1.00	0.97	1.10	1.09
Agreement	**H1-H2**		**H1-H2**	
Exact	60.6%		66.2%	
Adj.	37.4%		32.6%	
Non-A.	2.1%		1.2%	
Kappa	.446		.540	
QWK	.764		.844	
Correlation	.764		.844	

ENGINE TRAINING

The CRASE engine was trained using both of the human rater scores in the training sample. Separate models were trained for each of the ten marking criteria. The scoring models weighted the features differently for each criterion to maximize the agreement between the engine and the human rater scores. A set of five scoring models per marking criterion were developed during training so that the performance of the engine could be compared across different scenarios later on the cross-validation sample. The different models each produced continuous (i.e., non-integer scores) and cut points were set to place the continuous scores into categories that matched the marking guide for each marking criteria. Cuts points were chosen to ensure CRASE reasonably replicated the score distribution on the training sample.

ENGINE CROSS-VALIDATION

For each of the five scoring models, the cross-validation sample (sample 2, *n*=340) was scored by the engine using the models, parameters, and cuts generated during engine training. The integer scores (after cuts were applied) were used to evaluate model performance.

The performance of CRASE was evaluated relative to the human scores. Statistics used in the evaluation were: a) a comparison of score point distributions between the two scorers; b) a comparison of means and standard deviations between the two scorers; c) an examination of exact, adjacent, and non- adjacent agreement rates between the two scorers; d) an examination of correlations between the human and CRASE scores; and finally, e) an examination of the Kappa and Quadratic Weighted Kappa value. Quadratic Weighted Kappa is a measure of rater agreement that takes into consideration the agreement above and beyond chance and weights differences from exact agreement, with 'close' differences (e.g., adjacent scores) weighted more heavily.

Only the results of the final model are presented in this report. The model was selected from among the five models used by weighing multiple requirements: a) reasonably similar score point frequency distribution relative to both human rater scores; b) reasonably similar mean and standard deviation values relative to human rater scores; c) exact/adjacent/non-adjacent agreement rates that matched or exceeded that of the human raters, if possible; and, d) Quadratic Weighted Kappa values that matched or exceeded that of the human raters, if possible. Each of the five models performed reasonably well according to these requirements; however, the best-performing model was selected for this report and for providing scores to ACARA. As an example of a case where a model was not accepted, situations arose where the score point frequency distribution produced by the model did not represent the entire score range even though the agreement indices met requirements. We believed that the engine needed to, at a minimum, represent the score range even if it lowered agreement values.

The total scores across the marking criteria were not used in model selection, nor was a model built to specifically predict summed scores. The total scores were computed after the model was selected and were simply the sum of the marking criteria scores produced by the engine. This approach was taken because: a) we would expect that, if employed in operational assessment, the engine would be used to predict individual marking criterion scores; b) the quality of score prediction at the marking criteria level would be reasonably sufficient to ensure score prediction at the summed score level; and, c) we would expect the total score to be simply the sum of the marking criteria scores rather than produced by a separate model in an operational setting. The score frequency distributions, correlations, means, and standard deviations were computed on the summed scores.

Blind Evaluation Sample Scoring

Following the final model selection steps described in the previous section, the blind evaluation sample (sample 3, *n*=339) was submitted to the CRASE engine for scoring. The cuts, models, and parameters were employed based on the final model selected. Marking criterion score frequency distributions, means and standard deviations, and the summed score distributions, means, and standard deviations were then examined to evaluate whether they were similar to the training and cross-validation samples. We expected the distributional statistics to be similar across the samples, within sampling error. Because no human rater information was present for the blind sample, no further evaluation of model performance was possible.

Deliverables

A score file was produced and provided to ACARA that contained the scores for each of the samples using the final selected model. This file was simply the original file provided by ACARA but appended with 10 additional columns for the CRASE-supplied scores and one additional column for the summed CRASE-supplied scores across the ten marking criteria. The file was checked against the performance results to ensure that it produced the same statistics (distributional, agreement) for the training sample, the cross-validation sample, and the blind evaluation sample.

RESULTS

ENGINE TRAINING AND CROSS-VALIDATION SCORING RESULTS

This section of the report presents the performance of the final selected model on the training sample and the cross-validation sample. Score point distributional information and agreement statistics are presented for each marking criterion (Tables 7a-7j). Summed score distributional information and correlational data are provided as well (Tables 8 and 9). Across the ten marking criteria, the CRASE engine produced very similar mean and standard deviation scores relative to each of the human rater scores and produced similar score point distributions.

Table 6 presents a summary of the agreement results on the cross-validation sample across the ten marking criteria using the exact agreement, Kappa, and Quadratic Weighted Kappa statistics. In this table, the check mark (✓) is used to represent the status of CRASE agreement with human rater scores relative to the human-human agreement. The exact agreement, Kappa, and QWK statistics were calculated between the two human rater scores (H1-H2), between CRASE and the first human rater score (CRASE-H1), and between CRASE and the second human rater score (CRASE-H2). The agreement results were quite consistent across the three statistics. For seven of the marking criteria, both CRASE-human (CRASE-H1, CRASE-H2) agreements met or exceeded the human-human (H1-H2) agreements for the three statistics. For the "Paragraphing" criterion, neither of the CRASE-human agreement rates exceeded the human-human rates. For the "Punctuation" and "Cohesion" criteria, one (but not both) of the CRASE-human agreement rates exceeded the human-human agreement rates for each of the statistics.

Table 6. Summary of CRASE Agreement with First and Second Human Rater Scores and the Human-Human Agreement (Exact Agreement, Kappa, and QWK) on the Cross-Validation Sample

	Exact Agreement			Kappa			Quadratic Weighted Kappa		
Marking Criterion	**Both Exceed**	**One Exceed**	**None Exceed**	**Both Exceed**	**One Exceed**	**None Exceed**	**Both Exceed**	**One Exceed**	**None Exceed**
Audience	✓			✓			✓		
Text structure	✓			✓			✓		
Ideas	✓			✓			✓		
Persuasive devices	✓			✓			✓		
Vocabulary	✓			✓			✓		
Cohesion		✓			✓			✓	
Paragraphing			✓			✓			✓
Sentence structure	✓			✓			✓		
Punctuation		✓			✓			✓	
Spelling	✓			✓			✓		

Note. Both Exceed=CRASE-H1 and CRASE-H2 agreement values both exceed H1-H2 agreement value. One Exceed=One of CRASE-H1 or CRASE-H2 agreement values exceeds H1-H2 agreement value. None Exceed= None of CRASE-H1 or CRASE-H2 agreement values exceeds H1-H2 agreement value.

In addition to the summary information provided in Table 6, the performance of the final selected model of the CRASE engine is presented for the each marking criterion in Tables 7a-7j for both the training sample and the cross-validation sample. Because the training sample was used to build the engine models, there is some degree of overfit to these data as can be observed in the relative drop of the agreement statistics between the training sample and the cross-validation sample. The engine performances on the training sample are provided primarily for informational purposes. The cross-validation sample served as the evaluation sample of model performance because the model is applied to a new sample of student responses that were assumed to be drawn from the same population. Note that the C-H1 and C-H2 column headings in the Agreement sections of the table refer to the agreement between CRASE and each of the human raters (H1, H2).

Table 7a. Performance of Humans (H1, H2) and CRASE on the **"Audience" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)				Cross-Validation Sample (n=340)			
Score	**H1**	**H2**	**CRASE**		**H1**	**H2**	**CRASE**	
0	0.4%	0.4%	0.4%		0.0%	0.0%	0.0%	
1	3.0%	2.8%	2.5%		3.5%	2.4%	3.2%	
2	20.1%	18.0%	20.8%		20.3%	19.4%	20.6%	
3	35.9%	36.3%	33.7%		32.7%	35.0%	33.5%	
4	28.2%	26.0%	26.0%		27.7%	27.7%	27.4%	
5	8.7%	11.7%	12.0%		12.1%	11.5%	10.9%	
6	3.7%	4.7%	4.6%		3.8%	4.1%	4.4%	
Mean	3.29	3.39	3.36		3.36	3.39	3.35	
SD	1.11	1.16	1.17		1.15	1.12	1.15	
Agreement	**H1-H2**		**C-H1**	**C-H2**	**H1-H2**		**C-H1**	**C-H2**
Exact	62.6%		76.2%	77.1%	61.2%		64.4%	67.4%
Adj.	35.2%		23.3%	22.5%	36.5%		35.3%	31.2%
Non-Adj.	2.2%		0.4%	0.4%	2.4%		0.3%	1.5%
Kappa	.500		.683	.697	.485		.530	.566
QWK	.829		.904	.910	.821		.862	.855
Correlation	.833		.907	.911	.822		.862	.856

Table 7b. Performance of Humans (H1, H2) and CRASE on the **"Text Structure" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n =677)			Cross-Validation Sample (n=340)		
Score	**H1**	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
0	0.7%	0.9%	1.0%	1.5%	1.2%	1.2%
1	19.8%	18.9%	20.2%	20.0%	19.1%	19.4%
2	38.1%	37.1%	37.1%	37.1%	39.7%	37.4%
3	33.4%	31.6%	31.3%	32.9%	30.6%	33.5%
4	8.0%	11.5%	10.3%	8.5%	9.4%	8.5%
Mean	2.28	2.34	2.30	2.27	2.28	2.29
SD	0.90	0.94	0.94	0.93	0.92	0.91
Agreement	**H1-H2**		**C-H1** **C-H2**	**H1-H2**		**C-H1** **C-H2**
Exact	65.9%		80.2% 78.9%	68.8%		71.2% 69.7%
Adj.	33.7%		19.6% 21.1%	30.9%		28.8% 30.3%
Non-Adj.	0.4%		0.1% 0.0%	0.3%		0.0% 0.0%
Kappa	.517		.720 .704	.558		.591 .570
QWK	.790		.880 .881	.811		.830 .819
Correlation	.793		.881 .882	.811		.830 .819

Table 7c. Performance of Humans (H1, H2) and CRASE on the **"Ideas" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=676)			**Cross-Validation Sample (n=340)**		
Score	**H1**	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
0	0.7%	0.7%	1.0%	1.2%	0.6%	1.2%
1	4.6%	3.6%	3.6%	4.4%	3.5%	4.1%
2	20.4%	20.4%	21.1%	20.0%	23.5%	21.5%
3	49.9%	46.8%	46.2%	46.8%	44.1%	45.3%
4	18.8%	22.3%	22.2%	22.9%	22.1%	23.2%
5	5.5%	6.2%	5.9%	4.7%	6.2%	4.7%
Mean	2.98	3.05	3.03	3.00	3.02	2.99
SD	0.93	0.94	0.95	0.95	0.95	0.95

Agreement	**C-H1**	**C-H2**	**H1-H2**	**C-H1**	**C-H2**	**C-H1**
Exact	68.3%	79.4%	83.2%	62.4%	65.0%	69.4%
Adj.	31.1%	20.4%	16.8%	36.8%	34.7%	30.3%
Non-Adj.	0.6%	0.1%	0.0%	0.9%	0.3%	0.3%
Kappa	.533	.697	.755	.456	.491	.559
QWK	.809	.881	.906	.776	.801	.826
Correlation	.811	.883	.906	.776	.801	.826

Table 7d. Performance of Humans (H1, H2) and CRASE on the **"Persuasive Devices" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)			Cross-Validation Sample (n=340)		
Score	**H1**	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
0	0.9%	1.3%	1.2%	1.5%	1.5%	0.9%
1	17.7%	15.5%	16.1%	15.0%	17.4%	15.3%
2	40.8%	35.3%	38.4%	40.0%	35.3%	39.7%
3	31.5%	35.9%	28.5%	30.9%	33.8%	31.8%
4	9.2%	12.0%	15.8%	12.7%	12.1%	12.4%
Mean	2.30	2.42	2.42	2.38	2.38	2.39
SD	0.90	0.93	0.98	0.94	0.96	0.92
Agreement	**H1-H2**		**C-H1** **C-H2**	**H1-H2**		**C-H1** **C-H2**
Exact	59.8%		73.1% 76.5%	55.6%		66.5% 66.5%
Adj.	38.8%		26.7% 23.0%	42.9%		32.9% 33.5%
Non-Adj.	1.3%		0.1% 0.4%	1.5%		0.6% 0.0%
Kappa	.430		.622 .673	.377		.524 .528
QWK	.738		.845 .864	.727		.795 .809
Correlation	.745		.855 .865	.727		.795 .810

Table 7e. Performance of Humans (H1, H2) and CRASE on the **"Vocabulary" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)			Cross-Validation Sample (n=340)		
Score	**H1**	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
0	0.6%	0.7%	1.0%	0.9%	0.6%	1.2%
1	3.4%	2.4%	2.4%	2.9%	2.7%	2.4%
2	49.3%	46.2%	48.2%	47.9%	45.6%	48.2%
3	31.0%	29.3%	27.6%	29.1%	31.2%	28.8%
4	12.1%	15.6%	15.2%	14.4%	15.6%	14.7%
5	3.6%	5.5%	5.6%	4.7%	4.4%	4.7%
Mean	2.61	2.74	2.70	2.67	2.72	2.68
SD	0.89	0.97	0.98	0.95	0.93	0.96
Agreement	**H1-H2**		**C-H1** **C-H2**	**H1-H2**		**C-H1** **C-H2**
Exact	64.8%		77.0% 80.1%	64.4%		68.5% 73.2%
Adj.	33.4%		22.5% 19.2%	33.2%		30.3% 24.7%
Non-Adj.	1.8%		0.6% 0.7%	2.4%		1.2% 2.1%
Kappa	.467		.649 .702	.465		.524 .597
QWK	.768		.859 .883	.759		.807 .815
Correlation	.777		.867 .883	.760		.807 .816

Table 7f. Performance of Humans (H1, H2) and CRASE on the **"Cohesion" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)			Cross-Validation Sample (n=340)		
Score	**H1**	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
0	0.7%	0.7%	1.0%	0.9%	0.6%	1.2%
1	7.4%	8.3%	7.1%	5.6%	7.9%	5.6%
2	59.8%	56.1%	58.2%	58.8%	56.2%	57.4%
3	27.8%	28.1%	29.0%	30.0%	29.1%	32.4%
4	4.3%	6.8%	4.7%	4.7%	6.2%	3.5%
Mean	2.27	2.32	2.29	2.32	2.32	2.31
SD	0.69	0.75	0.71	0.69	0.73	0.69
Agreement	**H1-H2**		**C-H1** **C-H2**	**H1-H2**		**C-H1** **C-H2**
Exact	70.0%		80.6% 83.8%	71.8%		72.9% 69.1%
Adj.	28.7%		18.9% 16.2%	27.9%		27.1% 30.9%
Non-A.	1.3%		0.4% 0.0%	0.3%		0.0% 0.0%
Kappa	.481		.657 .721	.509		.517 .465
QWK	.674		.790 .848	.713		.714 .693
Correlation	.677		.790 .850	.714		.714 .694

Table 7g. Performance of Humans (H1, H2) and CRASE on the **"Paragraphing" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)				Cross-Validation Sample (n=340)			
Score	**H1**	**H2**	**CRASE**		**H1**	**H2**	**CRASE**	
0	21.1%	19.9%	22.6%		20.3%	20.6%	20.9%	
1	31.3%	31.6%	32.1%		34.1%	32.9%	33.5%	
2	32.1%	31.5%	28.2%		28.2%	32.4%	28.5%	
3	15.5%	17.0%	17.1%		17.4%	14.1%	17.1%	
Mean	1.42	1.45	1.40		1.43	1.40	1.42	
SD	0.99	0.99	1.02		1.00	0.97	1.00	
Agreement	**H1-H2**		**C-H1**	**C-H2**	**H1-H2**		**C-H1**	**C-H2**
Exact	65.9%		73.4%	75.2%	69.7%		65.9%	63.2%
Adj.	33.4%		26.3%	24.2%	29.7%		32.1%	34.7%
Non-Adj.	0.7%		0.3%	0.6%	0.6%		2.1%	2.1%
Kappa	.534		.638	.663	.585		.535	.497
QWK	.815		.863	.869	.834		.799	.778
Correlation	.816		.864	.870	.835		.799	.779

Table 7h. Performance of Humans (H1, H2) and CRASE on the **"Sentence Structure" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)				Cross-Validation Sample (n=340)			
Score	**H1**	**H2**	**CRASE**		**H1**	**H2**	**CRASE**	
0	0.9%	0.9%	1.0%		0.9%	0.6%	1.2%	
1	5.9%	4.3%	5.6%		5.3%	4.1%	4.7%	
2	24.5%	26.3%	28.1%		27.7%	27.7%	27.9%	
3	36.8%	34.9%	34.3%		33.8%	34.1%	34.4%	
4	24.4%	24.5%	23.0%		25.3%	24.4%	24.7%	
5	6.7%	6.9%	5.5%		5.6%	6.8%	5.6%	
6	0.9%	2.2%	2.5%		1.5%	2.4%	1.5%	
Mean	3.01	3.08	2.99		3.00	3.07	2.99	
SD	1.07	1.11	1.13		1.09	1.10	1.09	
Agreement	**H1-H2**		**C-H1**	**C-H2**	**H1-H2**		**C-H1**	**C-H2**
Exact	58.3%		71.3%	68.7%	58.5%		59.1%	59.4%
Adj.	38.3%		27.6%	30.9%	37.4%		38.8%	38.2%
Non-Adj.	3.4%		1.0%	0.4%	4.1%		2.1%	2.4%
Kappa	.437		.614	.579	.440		.446	.451
QWK	.780		.869	.870	.776		.800	.795
Correlation	.781		.870	.873	.777		.800	.797

Table 7i. Performance of Humans (H1, H2) and CRASE on the **"Punctuation" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)			Cross-Validation Sample (n=340)		
Score	**H1**	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
0	1.6%	2.1%	2.2%	1.5%	1.2%	1.8%
1	12.3%	10.9%	14.5%	10.9%	10.0%	9.7%
2	34.6%	31.8%	31.2%	35.0%	38.2%	36.5%
3	34.3%	35.0%	31.6%	35.3%	35.3%	36.8%
4	15.8%	17.7%	15.7%	15.0%	12.4%	12.4%
5	1.5%	2.5%	4.9%	2.4%	2.9%	2.9%
Mean	2.55	2.63	2.59	2.59	2.56	2.57
SD	1.00	1.04	1.13	1.00	0.97	0.99

Agreement	**H1-H2**	**C-H1**	**C-H2**	**H1-H2**	**C-H1**	**C-H2**
Exact	60.6%	64.8%	68.7%	60.6%	54.7%	63.2%
Adj.	35.6%	33.5%	30.9%	37.4%	42.6%	35.3%
Non-Adj.	3.8%	1.6%	0.4%	2.1%	2.6%	1.5%
Kappa	.458	.525	.580	.446	.364	.478
QWK	.755	.820	.861	.764	.728	.784
Correlation	.758	.827	.865	.764	.728	.784

Table 7j. Performance of Humans (H1, H2) and CRASE on the **"Spelling" Marking Criterion** on the Training and Cross-Validation Samples

	Training Sample (n=677)			Cross-Validation Sample (n=340)		
Score	**H1**	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
0	0.4%	0.6%	0.3%	0.9%	0.6%	0.6%
1	2.5%	1.9%	1.9%	1.8%	1.5%	2.1%
2	13.9%	11.7%	12.0%	10.9%	12.1%	10.9%
3	25.0%	26.0%	25.6%	22.7%	22.7%	22.9%
4	32.6%	32.5%	32.5%	38.2%	37.7%	38.8%
5	23.2%	25.6%	26.0%	23.5%	23.5%	22.9%
6	2.4%	1.8%	1.8%	2.1%	2.1%	1.8%
Mean	3.66	3.72	3.73	3.74	3.74	3.73
SD	1.15	1.11	1.10	1.10	1.09	1.08
Agreement	**H1-H2**	**C-H1**	**C-H2**	**H1-H2**	**C-H1**	**C-H2**
Exact	67.8%	77.3%	80.1%	66.2%	71.2%	69.4%
Adj.	30.7%	22.6%	19.8%	32.6%	28.2%	30.3%
Non-Adj.	1.5%	0.1%	0.1%	1.2%	0.6%	0.3%
Kappa	.572	.698	.733	.540	.607	.583
QWK	.854	.908	.917	.844	.872	.867
Correlation	.855	.911	.917	.844	.872	.867

Summary information about the summed scores across the ten marking criteria for each human rater score and for CRASE are presented in Table 8, along with the correlations between the human rater scores and CRASE (H1-H2, C-H1, C-H2). In the cross-validation sample, the CRASE-predicted summed score mean was similar to those of the two human rater scores. The CRASE summed score standard deviation was a bit larger than those provided by the human raters but the difference is not practically significant. The correlation between the two human rater scores and between CRASE and each human rater score were identical (.92) in the cross-validation sample.

Table 8. Means and Standard Deviations of Summed Scores across the Ten Marking Criteria (0-48) for Humans (H1, H2) and CRASE on the Training and Cross-Validation Samples

	Training Sample (n=677)			**Cross-Validation Sample (n=340)**		
	H1	**H2**	**CRASE**	**H1**	**H2**	**CRASE**
Mean	26.39	27.12	26.81	26.76	26.89	26.74
SD	8.52	8.91	9.37	8.76	8.70	9.13
Min	0	0	0	1	1	1
Max	48	48	48	48	48	48
	H1-H2	**C-H1**	**C-H2**	**H1-H2**	**C-H1**	**C-H2**
Correlation	.93	.96	.96	.92	.92	.92

Table 9 presents the frequency distributions of summed score for the human rater scores and CRASE. The data are very sparse for each scorer (CRASE, human) below summed score point 9 and each score point is represented by each scorer when the summed score is 9 or greater.

Table 9. Frequency Distribution of Summed Scores for Humans (H1, H2) and CRASE on the Training and Cross-Validation Samples

	Training Sample (n=677)						Cross-Validation Sample (n=340)					
Summed	H1		H2		CRASE		H1		H2		CRASE	
Score	N	%	N	%	N	%	N	%	N	%	N	%
0	2	0.3	2	0.3	1	0.2	0	0.0	0	0.0	0	0.0
1	2	0.3	3	0.4	2	0.3	3	0.9	2	0.6	1	0.3
2	0	0.0	0	0.0	3	0.4	0	0.0	0	0.0	3	0.9
3	0	0.0	0	0.0	1	0.2	0	0.0	0	0.0	0	0.0
4	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
5	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
6	0	0.0	1	0.2	0	0.0	0	0.0	1	0.3	0	0.0
7	3	0.4	3	0.4	2	0.3	1	0.3	0	0.0	0	0.0
8	2	0.3	0	0.0	1	0.2	1	0.3	0	0.0	1	0.3
9	7	1.0	7	1.0	5	0.7	3	0.9	3	0.9	2	0.6
10	4	0.6	1	0.2	5	0.7	2	0.6	1	0.3	4	1.2
11	3	0.4	3	0.4	4	0.6	1	0.3	3	0.9	4	1.2
12	4	0.6	4	0.6	8	1.2	2	0.6	1	0.3	2	0.6
13	11	1.6	10	1.5	8	1.2	5	1.5	5	1.5	3	0.9
14	11	1.6	13	1.9	13	1.9	5	1.5	4	1.2	2	0.6
15	17	2.5	11	1.6	23	3.4	8	2.4	9	2.7	7	2.1
16	24	3.6	18	2.7	18	2.7	9	2.7	12	3.5	16	4.7
17	16	2.4	11	1.6	19	2.8	7	2.1	9	2.7	10	2.9
18	13	1.9	26	3.8	19	2.8	10	2.9	8	2.4	10	2.9
19	18	2.7	24	3.6	23	3.4	10	2.9	13	3.8	10	2.9
20	26	3.8	19	2.8	20	3.0	10	2.9	10	2.9	11	3.2
21	20	3.0	22	3.3	20	3.0	13	3.8	13	3.8	9	2.7
22	31	4.6	27	4.0	31	4.6	19	5.6	8	2.4	17	5.0
23	39	5.8	32	4.7	34	5.0	18	5.3	18	5.3	12	3.5
24	39	5.8	38	5.6	33	4.9	20	5.9	18	5.3	27	7.9
25	35	5.2	37	5.5	40	5.9	14	4.1	17	5.0	24	7.1
26	23	3.4	28	4.1	25	3.7	14	4.1	18	5.3	11	3.2
27	32	4.7	30	4.4	25	3.7	15	4.4	13	3.8	2	0.6
28	25	3.7	23	3.4	17	2.5	17	5.0	11	3.2	6	1.8
29	29	4.3	23	3.4	24	3.6	9	2.7	14	4.1	12	3.5
30	32	4.7	23	3.4	18	2.7	8	2.4	20	5.9	13	3.8
31	26	3.8	23	3.4	21	3.1	10	2.9	11	3.2	13	3.8

Summed Score	Training Sample (n=677)						Cross-Validation Sample (n=340)					
	H1		H2		CRASE		H1		H2		CRASE	
	N	%	N	%	N	%	N	%	N	%	N	%
32	26	3.8	23	3.4	27	4.0	16	4.7	12	3.5	14	4.1
33	22	3.3	26	3.8	19	2.8	11	3.2	10	2.9	13	3.8
34	18	2.7	24	3.6	14	2.1	10	2.9	14	4.1	12	3.5
35	25	3.7	22	3.3	17	2.5	12	3.5	11	3.2	6	1.8
36	12	1.8	18	2.7	24	3.6	9	2.7	6	1.8	10	2.9
37	15	2.2	13	1.9	13	1.9	8	2.4	3	0.9	9	2.7
38	3	0.4	12	1.8	11	1.6	8	2.4	7	2.1	8	2.4
39	16	2.4	13	1.9	22	3.3	6	1.8	4	1.2	10	2.9
40	7	1.0	10	1.5	8	1.2	2	0.6	8	2.4	2	0.6
41	6	0.9	9	1.3	11	1.6	5	1.5	3	0.9	6	1.8
42	7	1.0	10	1.5	12	1.8	4	1.2	3	0.9	1	0.3
43	4	0.6	7	1.0	7	1.0	3	0.9	3	0.9	3	0.9
44	6	0.9	5	0.7	3	0.4	3	0.9	3	0.9	2	0.6
45	6	0.9	6	0.9	6	0.9	5	1.5	4	1.2	3	0.9
46	5	0.7	10	1.5	7	1.0	1	0.3	1	0.3	4	1.2
47	4	0.6	6	0.9	10	1.5	2	0.6	3	0.9	1	0.3
48	1	0.2	1	0.2	3	0.4	1	0.3	3	0.9	4	1.2

Figure 1 presents the histogram of the summed scores for each of the human raters and CRASE for the training sample. Figure 2 presents a visual depiction of the cumulative frequency distribution (CFD) of the summed scores for each of the human raters and for CRASE for the training sample. Both figures suggest that the three scoring distributions are quite similar with Human 1 differing slightly from the Human 2 and CRASE scores at or about summed score 32.

Figure 1. Histogram of Summed Scores for Humans (Human 1, Human 2) and CRASE on the **Training Sample**

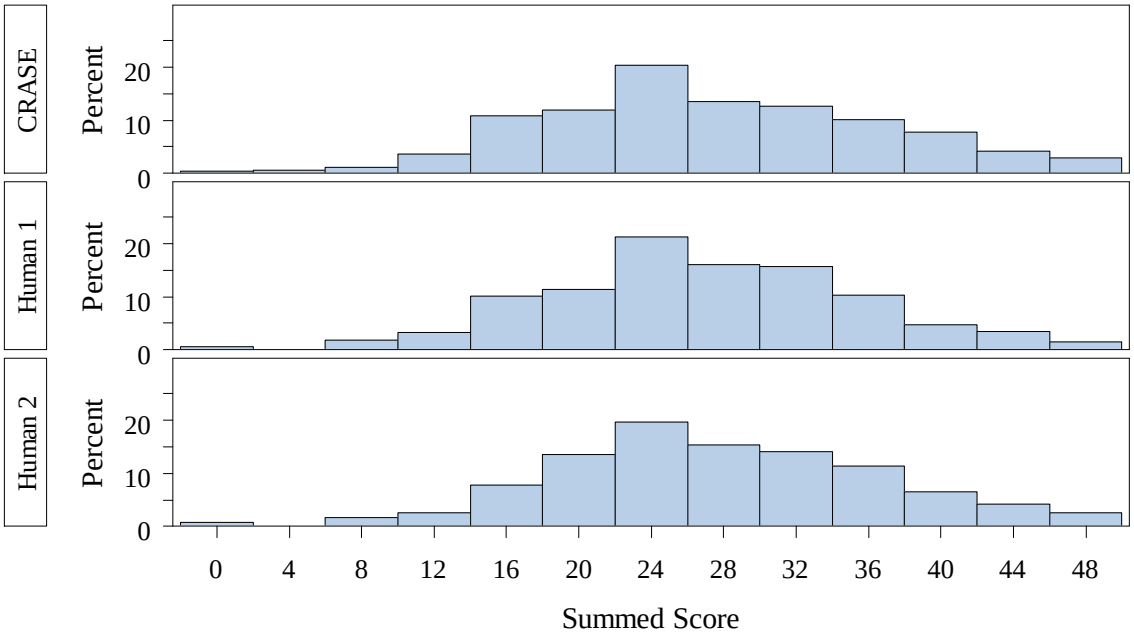


Figure 2. Cumulative Frequency Distribution of Summed Scores for Humans (Human 1, Human 2) and CRASE on the **Training Sample**

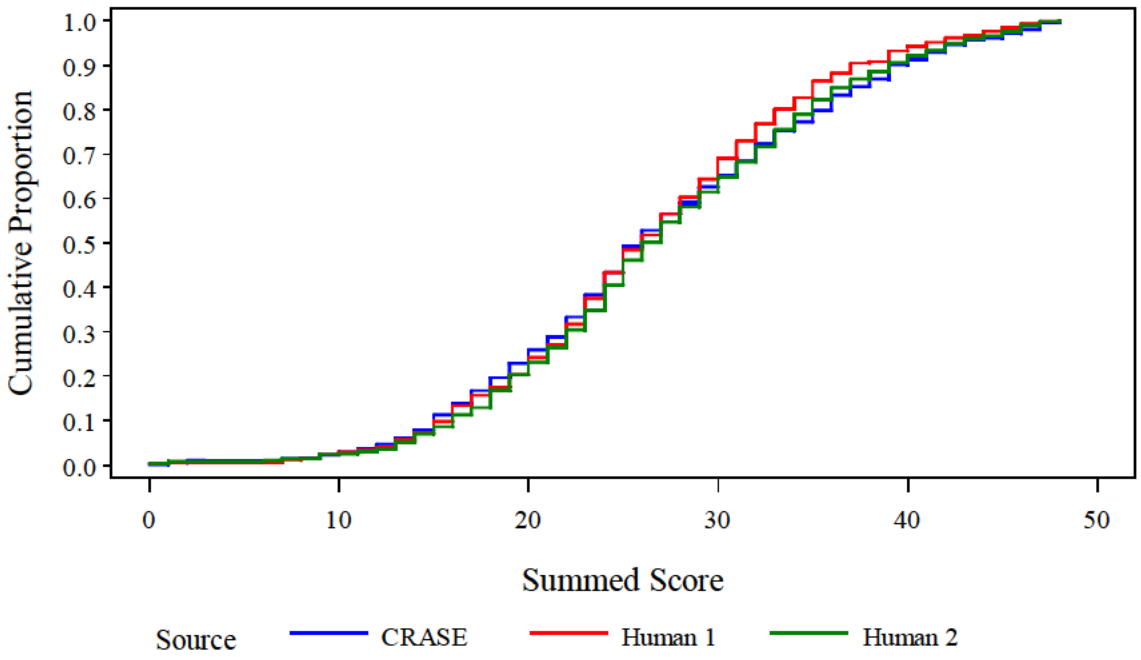


Figure 3 presents the histogram of the summed scores for each of the human raters and CRASE for the cross-validation sample. Figure 4 presents a visual depiction of the cumulative frequency distribution of the summed scores for each of the human raters and for CRASE for the cross-validation sample. Both Figures suggest that the three scoring distributions are quite similar, although CRASE assigns fewer summed scores around score point 28 and higher summed scores around score point 32.

Figure 3. Histogram of Summed Scores for Humans (Human 1, Human 2) and CRASE on the **Cross-Validation Sample**

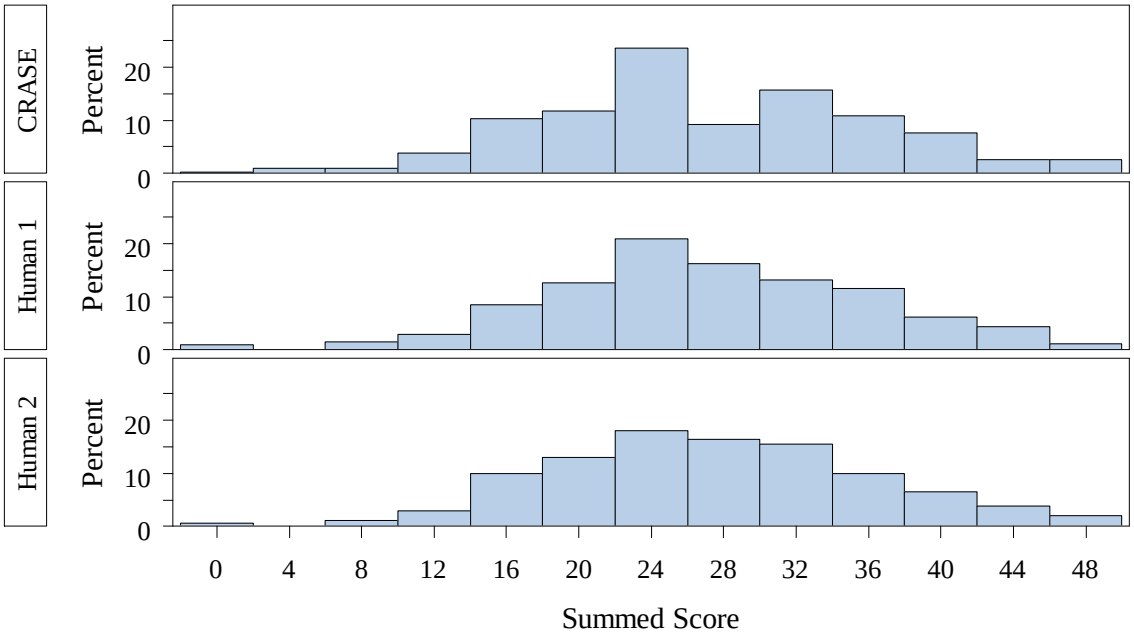
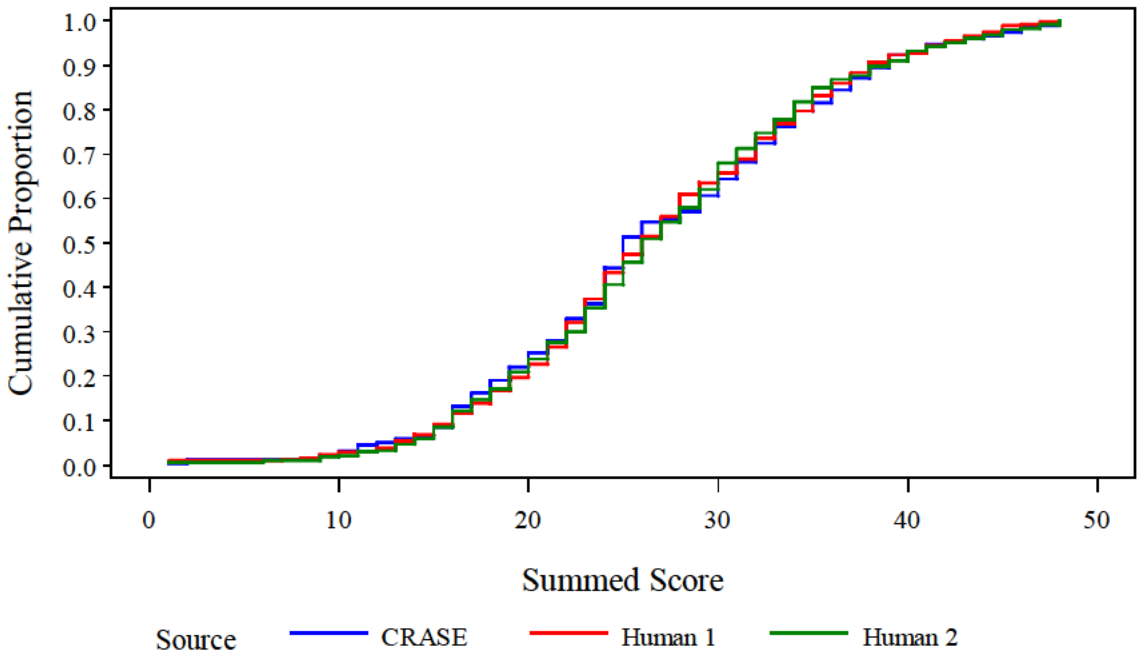


Figure 4. Cumulative Frequency Distribution of Summed Scores for Humans (Human 1, Human 2) and CRASE on the **Cross-Validation Sample**



ENGINE BLIND EVALUATION SAMPLE SCORING RESULTS

The CRASE-assigned results on the blind evaluation sample are presented in Tables 10-12. Because this sample serves as a blind evaluation sample for ACARA, the human-assigned scores were not provided to Pacific Metrics for these responses. As a result, only CRASE-assigned scores are presented. The score distributional information for the ten marking criteria and the summed scores are similar to those appearing in the training and cross-validation samples.

Table 10. Score Point Distributions, Means, and Standard Deviations of CRASE-predicted Scores on the 10 Marking Criterion the Blind Evaluation Sample (*n*=339)

Score	Audience	Text Structure	Ideas	Persuasive Devices	Vocabulary
0	0.3%	1.5%	1.5%	1.8%	1.5%
1	3.0%	21.5%	4.1%	16.8%	2.7%
2	22.4%	34.2%	22.1%	35.1%	45.7%
3	29.2%	33.0%	44.3%	32.2%	31.9%
4	29.8%	9.7%	24.8%	14.2%	14.8%
5	13.0%	.	3.2%	.	3.5%
6	2.4%	.	.	.	.
Mean	3.34	2.28	2.96	2.40	2.66
SD	1.14	0.96	0.95	0.98	0.94
Score	**Cohesion**	**Paragraphing**	**Sentence Structure**	**Punctuation**	**Spelling**
0	1.5%	21.5%	1.5%	2.7%	0.6%
1	5.9%	31.9%	5.3%	12.7%	2.1%
2	55.8%	31.0%	26.6%	32.2%	11.2%
3	35.1%	15.6%	34.2%	33.6%	23.3%
4	1.8%	.	27.4%	15.9%	37.8%
5	.	.	4.4%	3.0%	24.2%
6	.	.	0.6%	.	0.9%
Mean	2.30	1.41	2.96	2.56	3.72
SD	0.67	0.99	1.06	1.08	1.08

The summed score mean and standard deviation is also similar to those produced by CRASE and the human rater scores in the training and cross-validation samples (Table 11).

Table 11. Means and Standard Deviations of CRASE-derived Summed Scores (0-48) on the Blind Evaluation Sample (*n*=339)

Statistic	CRASE
Mean	26.60
SD	9.22

Table 12 presents the frequency distribution of the summed scores across the ten marking criteria in the blind evaluation sample.

Table 12. Frequency Distribution of Summed Scores from CRASE on Blind Evaluation Sample (*n*=339)

Summed Score	N	%	Summed Score	N	%
0	0	0.0	24	20	5.9
1	2	0.6	25	25	7.4
2	2	0.6	26	4	1.2
3	0	0.0	27	7	2.1
4	1	0.3	28	5	1.5
5	0	0.0	29	14	4.1
6	0	0.0	30	16	4.7
7	1	0.3	31	12	3.5
8	1	0.3	32	16	4.7
9	1	0.3	33	13	3.8
10	5	1.5	34	13	3.8
11	1	0.3	35	10	3.0
12	5	1.5	36	3	0.9
13	3	0.9	37	3	0.9
14	6	1.8	38	12	3.5
15	8	2.4	39	11	3.2
16	18	5.3	40	11	3.2
17	8	2.4	41	4	1.2
18	8	2.4	42	2	0.6
19	10	3.0	43	4	1.2
20	9	2.7	44	3	0.9
21	8	2.4	46	1	0.3
22	15	4.4	47	3	0.9
23	14	4.1	48	1	0.3

Figure 5 presents the histogram of the summed scores CRASE for the blind evaluation sample. Figure 6 presents a visual depiction of the cumulative frequency distribution of the summed scores for CRASE for the blind evaluation sample.

Figure 5. Histogram of Summed Scores from CRASE on **Blind Evaluation Sample** (*n*=339)

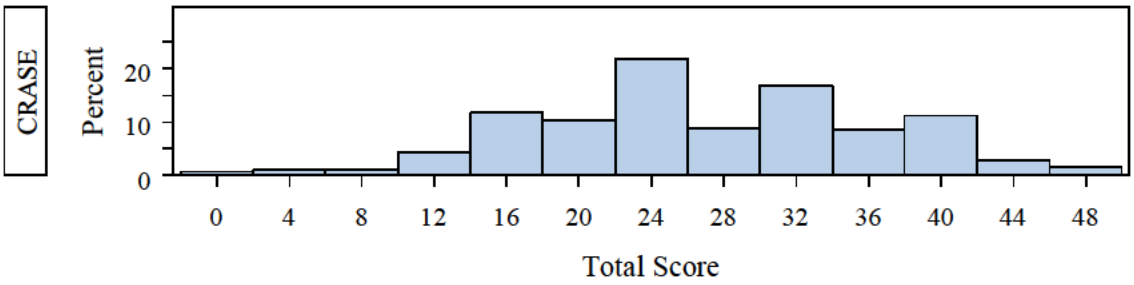


Figure 6. Cumulative Frequency Distribution of Summed Scores for CRASE on the **Blind Evaluation Sample** (*n*=339)

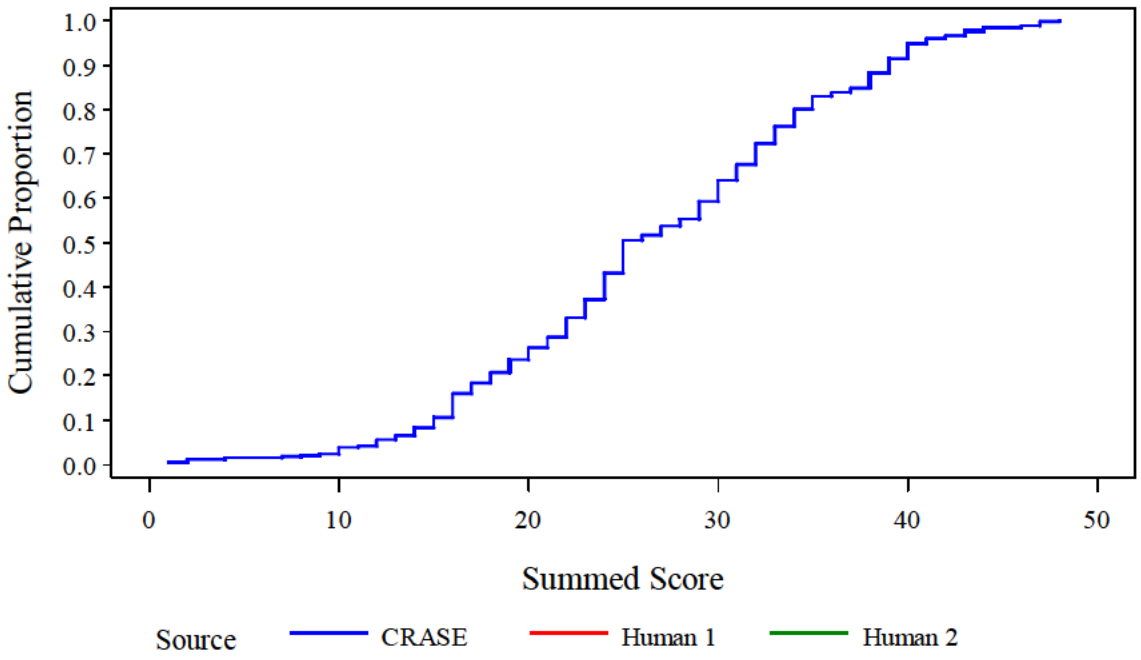


Figure 7 presents the histogram of the CRASE-derived summed scores for all three samples (training, cross-validation, blind evaluation). This figure suggests that the cross-validation and blind evaluation summed scores assigned by CRASE are quite similar and slightly different than training sample summed scores in the upper part of the score range. Figure 8 presents a visual depiction of the cumulative frequency distribution of the summed scores for CRASE for all three samples. This figure suggests that CRASE produces similar CDFs for each of the samples.

Figure 7. Histogram of Summed Scores from CRASE on All Three Samples

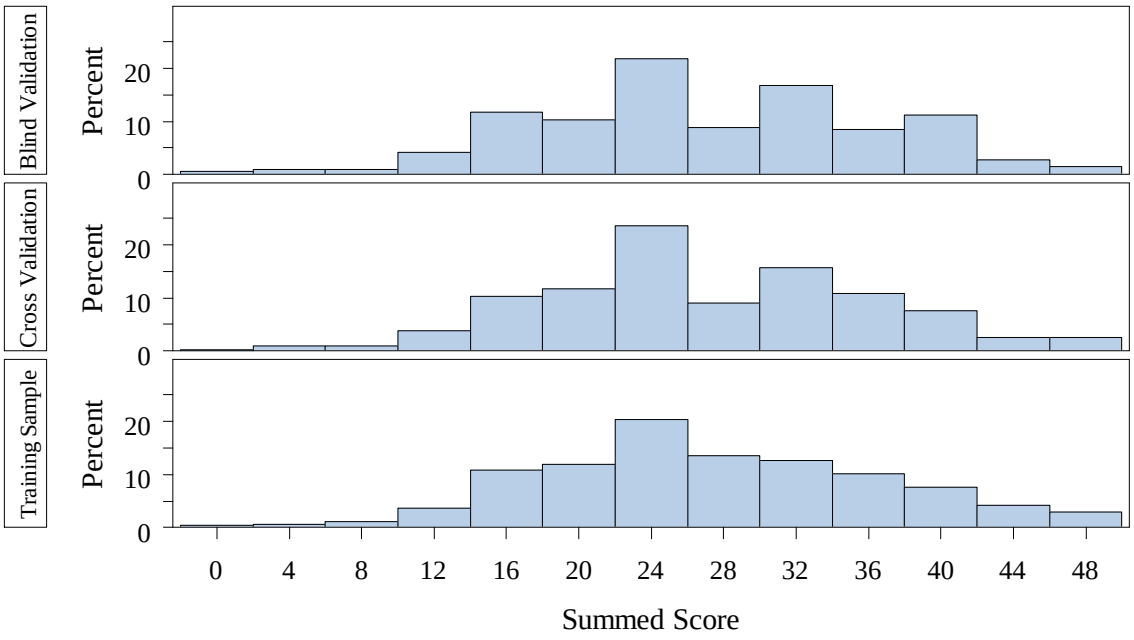
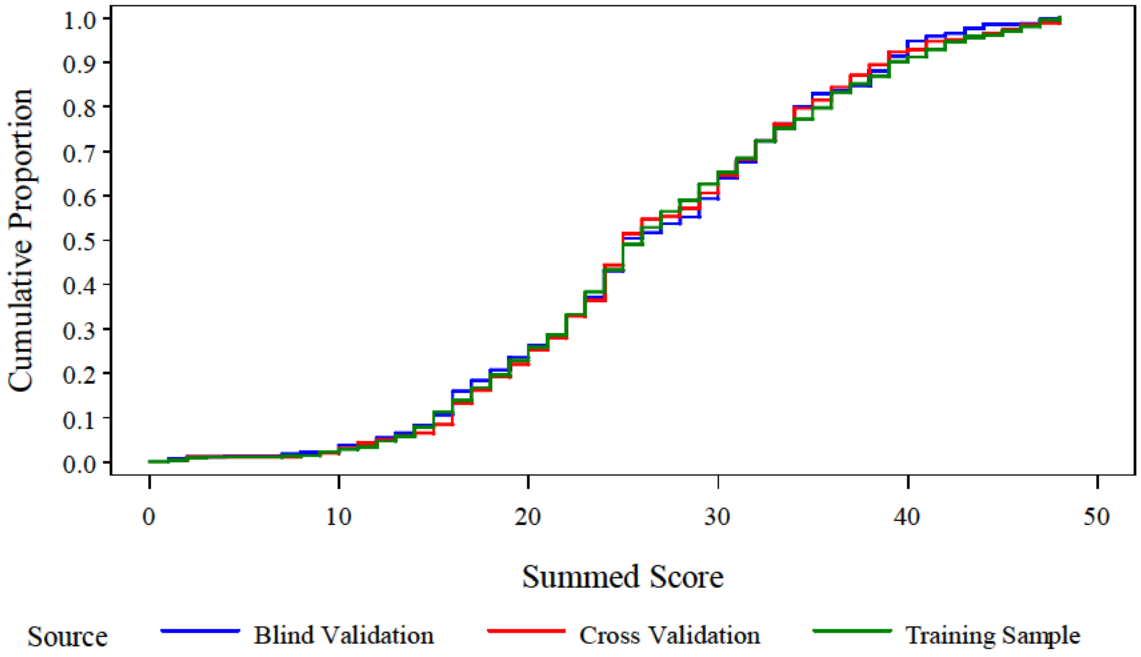


Figure 8. Cumulative Frequency Distribution of Summed Scores for CRASE on All Three Samples



SUMMARY AND CONCLUSION

In conclusion, the results of the study on the cross-validation sample suggest that the CRASE engine performs very well relative to the human raters, with CRASE-human agreement often exceeding human-human agreement. Across the ten marking criteria, the CRASE engine produced very similar mean and standard deviation scores relative to each of the human rater scores and produced similar score point distributions. For seven of the marking criteria, both CRASE-human (CRASE-H1, CRASE-H2) agreements met or exceeded the human-human (H1-H2) agreements for the three statistics. For the "Paragraphing" criterion, neither of the CRASE-human agreement rates exceeded the human-human rates. For the "Punctuation" and "Cohesion" criteria, one of the CRASE-human agreement rates exceeded the human-human agreement rates for each of the statistics.

NAPLAN Online Trial Study 2013
Automated Essay Scoring of Writing Scripts

Scoring with the MetaMetrics Auto-Essay Scoring Engine and The Lexile® Framework for Writing

29 August 2013

Prepared by MetaMetrics for:

Australian Curriculum, Assessment and Reporting Authority
Level 10, 255 Pitt Street,
Sydney, New South Wales, Australia 2000

MetaMetrics.

1000 Park Forty Plaza Drive, Suite 120
Durham, North Carolina 27713
www.MetaMetricsInc.com
www.Lexile.com
www.Quantiles.com

ACARA NAPLAN
Online Trial Study 2013 – Automated Essay Scoring Employing The Lexile® Framework for Writing

MetaMetrics, an educational measurement and research organization, was pleased to respond to the Invitation to Quote released by the Australian Curriculum, Assessment and Reporting Authority (ACARA) in relation to the NAPLAN Online Trial Study 2013 – Automated Essay Scoring of Writing Scripts and Report.

The Lexile® Framework for Reading had its developmental roots in the mid-80s and has been refined ever since as a scale to measure text complexity and, more importantly, place students' reading ability levels on that same scale. It is the only metric that produces this unique relationship, enabling teachers, parents, and students to select instructional materials appropriately targeted to the reader. With the introduction of The Lexile Framework for Writing in 2007, MetaMetrics has moved a step closer to its goal of measuring the four modalities of communication within a common framework and reporting methodology.

Reading Next (2004) and *Writing Next* (2007) have documented the importance of the reading-writing connection. Both reports affirm that students' reading and writing abilities are complimentary and growth in one skill inevitably leads to growth in the other (i.e., students become better readers by strengthening their writing skills and vice-versa). The Lexile Framework for Writing expresses student writing ability on the same Lexile scale as reading ability. This approach provides educators with a consistent and straight forward method to measure and monitor student growth in both reading and writing and reinforces the importance of reading in the development of writing skills.

About MetaMetrics

MetaMetrics, an educational measurement and research organization based in Durham, North Carolina, is dedicated to "Bringing Meaning to Measurement." The genesis of the organization was predicated upon the notion that assessment and instruction could and should be connected. Our founders, Dr. A. Jackson Stenner and Dr. Malbert Smith, had a vision to make test scores more actionable by blurring the distinction between assessment and instruction. With this corporate vision, MetaMetrics was created in 1984, built upon the optimism and passion of two individuals who thought they could make a positive contribution to educating all students.

This vision of the future was shared by scientists at The National Institute of Child Health and Human Development (NICHD) who funded MetaMetrics' research with a series of grants over the course of a decade. These grants supported research on reading and psychometric theory, which culminated in the development of MetaMetrics' flagship product, The Lexile Framework for Reading (www.Lexile.com).

The creation of The Lexile Framework for Reading marks the first attempt in education to unify the measurement of reading. Dr. Stenner and Dr. Smith believed that one of the major impediments to progress in the social (soft) sciences versus the hard sciences was in the proliferation of tests and measurement systems. What philosophers of science call the *unification of measurement* was absent in education. With the creation of the Lexile Framework, Stenner and Smith demonstrated that common scales, like Fahrenheit and Celsius, could be built for reading.

MetaMetrics develops scientific measures of student achievement and complementary technologies that link assessment with targeted instruction to improve learning. The organization employs a highly skilled staff with diverse backgrounds. MetaMetrics' staff has over 70 years of work experience in state education agencies, and over 200 years of teaching experience at the elementary through university level. The organization's staff holds more than 40 doctorate and graduate degrees from several of the most prestigious universities in the world, including Duke University, Princeton University, Stanford University, and The University of North Carolina at Chapel Hill.

MetaMetrics' renowned team of psychometricians have completed over 30 linking studies to state assessments, participated in three national studies for the National Center for Educational Statistics (NCES), and have developed more than 20 interim assessments. The team of psychometricians and their research agenda are supported by the founders, Dr. Stenner and Dr. Smith, who continue to publish and present papers at major international and national assessment conferences. Both Dr. Stenner and Dr. Smith hold joint appointments as research professors at the School of Education at the University of North Carolina. Supporting the research and development team are two senior scientists, Dr. Donald Burdick (Professor Emeritus, Duke University) and Dr. Jill Fitzgerald, who was recently inducted into the International Reading Association (IRA) Hall of Fame.

For nearly 30 years, MetaMetrics' work has been recognized worldwide for its distinct value in differentiating instruction and personalizing learning. Its products and services for reading (The Lexile Framework for Reading, El Sistema Lexile® para Leer), writing (The Lexile Framework for Writing), and mathematics (The Quantile® Framework for Mathematics) are utilized throughout the world. Built upon these foundations, MetaMetrics also has created personalized learning platforms (Engaging English®, Learning Oasis™) to facilitate the growth of reading and writing ability.

MetaMetrics has completed numerous linking studies related to reading, writing, and mathematics for state education agencies, national research institutes, and assessment publishers. Each linking study is conducted according to a specific schedule that is developed with the partner and is coordinated with the administration of their high-stakes assessment. Validation studies are conducted to support the inferences made by partners and states when using the Lexile and Quantile metrics.

MetaMetrics is the sole source for the reading, writing, and mathematics measurement system frameworks utilizing Lexile and Quantile measures. The scientific algorithms that produce a Lexile measure and a Quantile measure for both text/resource/essay and student are proprietary to and available only from MetaMetrics.

About The Lexile Framework for Writing

The Lexile Framework for Writing is a scientific approach to measuring writing ability, utilizing the same Lexile scale used to measure reading ability and text complexity. A common, developmental scale provides educators with a consistent and straightforward method for monitoring student growth in both reading and writing and reinforces the importance of reading in the development of writing skills (ranges from "BW" – Beginning Writer to greater than 1500W). A Lexile writer measure is an estimate of a student's ability to express language in writing, based on factors related to semantic complexity (the level of words used) and syntactic sophistication (how the words are combined into sentences). A student's Lexile writer measure typically is lower than his or her Lexile reader measure, as students tend to comprehend text at a higher Lexile level than they can produce written text. Because the writing scale is grade-, genre-, prompt-, and punctuation-independent, educators can use students' Lexile writer measures to differentiate instruction and monitor the development of writing skills over time and across the curriculum (grades 2–12). All Lexile writing products and services rely on the Lexile writer measure and Lexile scale to match a writer with targeted writing materials and activities (e.g., practice with increasingly sophisticated writing conventions and devices).

A Lexile writer measure refers to an underlying individual trait, which is defined as the power to compose written text, with writing ability embedded in a complex web of cognitive and sociocultural processes. Individuals with higher-level writing ability are more facile with at least some of the aspects of a writer-composition-reader transaction than are individuals with lower-level writing ability. Facets of a writer-composition-reader transaction may be related to, reflected in, or reflective of, an individual's writing ability, but they are not, in themselves writing ability. Rather, writing ability is an individual trait that is brought to bear to greater or lesser extent within each transaction occasion.

The Lexile Writing Analyzer is a grade-, genre-, prompt-, and punctuation-independent automatic essay-scoring engine for establishing Lexile writer measures. Through a research study to examine the relationship between text complexity, text features, and writing ability, a parsimonious set of significant predictors emerged—predictors consistent with the hypothesis that selected kinds of composition surface text features may be proxies for degree of executive functioning and working memory capacity and efficiency (Burdick, Swartz, Stenner, Fitzgerald, Burdick, and Hanlon, in press). The resulting combination consisted of lexical representations alone—without syntax signifiers. Specifically, a combination of a small number of variables—degree of diverse use of vocabulary and greater vocabulary density, controlling for production fluency—predicted 90% of the true variance in rater judgments of essays. The correlation between the average of four human raters and the Lexile writer measures for a sample of 663 students was .78 [disattenuated correlation was .95] (Swartz, Burdick, Stenner, Burdick, Hanlon, and Hooper, 2010).

Measurement precision of the Lexile writer measures was addressed by examining the average standard error of measurement and generalizability coefficients in relation to the increasing number of essays per student (Burdick, Swartz, Stenner, Fitzgerald, Burdick, and Hanlon, in press). The generalizability coefficients were high and very similar across 1 to 6 essays for human raters (.71 for one essay to .94 for six essays) and Lexile writer measures (.70 for one essay to .93 for six essays).

In 2008, MetaMetrics conducted a validity study for the North Carolina Department of Public Instruction (NCDPI). The purpose of this research was to investigate the potential for linking the writing scoring model used by the North Carolina Department of Public Instruction with The Lexile Framework for

Writing. Estimating writer ability using the Lexile Writing Analyzer was the focus of the research reported in this document. NCDPI provided MetaMetrics with 3,032 handwritten responses with Total Writing Scores (range of scores was 4 to 20) from students in grades 4 (*N* = 1,030), 7 (*N* = 991), and 10 (*N* = 1,011). The handwritten responses were transcribed and then analyzed using the Lexile Writing Analyzer, an automatic essay scoring engine. The results suggest that the holistic scoring model used in North Carolina is a valid indicator of writing quality. The results showed a monotonically increasing function across grades of the Lexile writer measures (Grade 4, mean = 759W, SD = 435W; Grade 7, mean = 986W, SD = 442W; and Grade 10, mean = 1392W, SD = 492W). The correlations between Total Writing Score and Lexile writer measure ranged from .73 (Grade 10) to .83 (Grade 4). While the holistic scoring model does not communicate information about growth underlying writer ability (i.e., the use of words and how words are combined), when combined with the results from The Lexile Framework for Writing a more complete picture of a writer is revealed.

In 2012, MetaMetrics participated in a study to compare the results from nine automated essay scoring engines on eight essay scoring prompts (Shermis and Hammer). A total of 22,029 essays were scored and the "…results demonstrated that overall, automated essay scoring was capable of producing scores similar to human scores for extended-response writing items with equal performance for both source-based and traditional writing genre" (p. 2). The quadratic-weight kappa statistic was used to evaluate the relationship between the human scores and the machine scores. For the Lexile Writing Analyzer, the quadratic-weighted kappa statistics for the eight essay sets ranged from .55 to .68 (mean .65). While typically lower than the quadratic-weighted kappa statistics between the two human ratings (typical range .66 to .85 for human rater performance in statewide high-stakes testing programs), the Lexile Writing Analyzer was higher on one set of essays.

Auto-Essay Scoring Feasibility Study

The National Assessment Program — Literacy and Numeracy (NAPLAN) is the main assessment program conducted by ACARA. Every year over one million Australian students in Years 3, 5, 7, and 9 are assessed in Reading, Writing, Language Conventions, and Numeracy. The tests are equated longitudinally each year in order to ensure comparability of results from year to year and to enable tracking of students' results. The tests were first administered in 2008.

Between September and December 2012, ACARA conducted a pilot research study which investigated how the change of test delivery mode (from paper to computer) would affect student writing performance. The essays to be scored were written by students from Years 3, 5, 7, and 9 who participated in this pilot study. Students completed the writing tasks on computers online. Four hundred essays at each year level were included in this feasibility study to be scored automatically.

NAPLAN Writing Test -- Prompt and Rubric

The NAPLAN writing test task "targets the full range of student capabilities expected of students from Years 3 to 9" (ACARA, 2013). A single prompt (stimulus) is administered to all students in Years 3, 5, 7, and 9 and the same scoring rubric is used with all papers with the expectation that "more capable writers will address the topic at a higher level." For this pilot research study, students completed a persuasive writing task.

In the *2012 NAPLAN Persuasive Writing Marking Guide* (ACARA, 2012), "the purpose of persuasive writing is to persuade a reader to a point of view on an issue. Persuasive writing may express an opinion, discuss, analyze and evaluate an issue. It may also entertain and inform. The style of persuasive writing may be formal or informal but it requires the writer to adopt a sense of authority on the subject matter and to develop the subject in an ordered, rational way. A writer of a persuasive text may draw on their own personal knowledge and experience or may draw on detailed knowledge of a particular subject or issue. The main structural components of the persuasive text are the introduction, development of argument (body) and conclusion" (page 5). Student essays are rated using an analytic, criterion-referenced marking guide and each persuasive essay is marked for the following ten criteria:

- *Audience* (range of scores 0-6). The writer's capacity to orient, engage, and persuade the reader.
- *Text structure* (range of scores 0-4). The organization of the structural components of a persuasive text (introduction, body, and conclusion) into an appropriate and effective text structure.
- *Ideas* (range of scores 0-5). The selection, relevance, and elaboration of ideas for a persuasive argument.
- *Persuasive devices* (range of scores 0-4). The use of a range of persuasive devices to enhance the writer's position and persuade the reader.
- *Vocabulary* (range of scores 0-5). The range and precision of contextually appropriate language choices.
- *Cohesion* (range of scores 0-4). The control of multiple threads and relationships across the text, achieved through the use of grammatical elements (referring words, text connectives, conjunctions) and lexical elements (substitutions, repetitions, word associations).

- *Paragraphing* (range of scores 0-3). The segmenting of text into paragraphs that assists the reader to follow the line of argument.
- *Sentence structure* (range of scores 0-6). The production of grammatically correct, structurally sound, and meaningful sentences.
- *Punctuation* (range of scores 0-5). The use of correct and appropriate punctuation to aid the reading of the text.
- *Spelling* (range of scores 0-6). The accuracy of spelling and the difficulty of the words used.

NAPLAN Writing Pilot Study

Students are expected to produce an extended essay style response of between 200 and 800 words in length. During the 2012 pilot study administration, students accessed the writing test and responded online through an online test delivery system provided by ACARA. All essays were marked by two raters for the ten (10) criteria scores and the Total Score.

Sample

ACARA provided a set of 1,356 essays from the 2012 pilot research study administration and divided the set of essays into three samples. Sample 1 consisted of approximately 50% of the essays (*N* = 677, 49.9%) and the purpose was to conduct training of the MetaMetrics auto-essay scoring (AES) engine. Sample 2 consisted of approximately 25% of the essays (*N* = 340, 25.1%) and the purpose was for validation of the algorithm developed during the training of the AES engine with the essays from Sample 1. For Samples 1 and 2, all ratings were provided from the two raters for the ten criteria and the Total score. Sample 3 consisted of approximately 25% of the essays (*N* = 339, 25.0%) and the purpose was for independent validation of the results from the MetaMetrics AES engine trained with the essays from Samples 1 and 2.

Essays were examined for non-valid essays (copied prompt, rating outside criteria range, or no score by raters). Three essays were not measured because of invalid data (all from Sample 1).

- 1325830 -- It appears that the student was just trying to copy the prompt.
- 1326080 -- This essay contains a rating of 9 for rating category 3 which is outside the range in the documentation.
- 1326570 -- No scores by raters and an LWA of -10000.

Table 1 presents the descriptive statistics for the NAPLAN Writing Test Total Score for the three samples identified by ACARA. The results show that the three samples are equivalent in terms of Total Score.

Table 1. Descriptive statistics for essay Total scores (average), by sample.

Sample	*N*	Total Score Rater 1 Mean (SD)	Total Score Rater 2 Mean (SD)	Total Score Average Mean (SD)
1	674	26.47 (8.42)	27.21 (8.81)	26.84 (8.45)
2	340	26.76 (8.76)	26.89 (8.70)	26.82 (8.54)
1 & 2	1,014	26.57 (8.53)	27.10 (8.77)	26.83 (8.47)

Across Samples 1 & 2 combined, the correlation between the two raters was 0.92 ($p < .0001$). The correlations of the criteria scores for the two raters ranged from 0.68 (cohesion) to 0.85 (spelling).

To examine the relationship between the Total Score and the ten marking criteria, the correlations between the scores are presented in *Table 2* for each rater (data from Sample 1 & 2). All of the correlations are significant at the .0001 level indicating that the Total Score is measuring the same construct as the ten marking criteria.

Table 2. Correlations between Total Score and marking criteria (Samples 1 & 2, *N* = 1,014), by rater.

Marking Criteria	Rater 1 Correlation with Total Score	Rater 2 Correlation with Total Score
Audience	0.94	0.94
Text Structure	0.90	0.90
Ideas	0.91	0.92
Persuasive Devices	0.89	0.91
Vocabulary	0.88	0.90
Cohesion	0.85	0.87
Paragraphing	0.87	0.88
Sentence Structure	0.91	0.91
Punctuation	0.79	0.80
Spelling	0.89	0.89

Additionally, the intercorrelations of the marking criteria range from 0.621 (Paragraphing with Punctuation for Rater 1) to 0.878 (Audience with Ideas for Rater 2), with most in the upper 70s.

Scores and Results

Lexile Essay Score. MetaMetrics generally provides both research scores and "reported" scores for all of our work (Lexile measures for reading and writing and Quantile measures). The reason for this is because of the extent of measurement error at the tails of the distributions and, therefore, the decreased instructional relevance. For Lexile essay scores, MetaMetrics' provides two "caps":

- At the lower end, all scores of 0W (which stands for 0 Lexile in writing) and below are reported as "BW" for "Beginning Writing"; and
- At the upper end, all scores of 2000W and above are capped at 2000W.

For this study, only research Lexile essay scores were provided. Essays with less than 50 words are typically flagged indicating the measures may have increased measurement error due to the small sample of data. *Table 3* presents the Lexile essay score summary statistics for the various samples. As with the NAPLAN Total Scores, the samples are comparable when measured with the Lexile Writing Analyzer.

Table 3. Descriptive statistics for Lexile essay scores, by sample.

Sample	*N*	Lexile Essay Score Mean (SD)	Lexile Essay Score Range
1	674	719.89 (426.61)	-249W to 2000W
2	340	720.54 (419.70)	-116W to 2000W
1 & 2	1,014	720.10 (424.10)	-249W to 2000W
3	339	724.75 (408.45)	-309W to 1932W
1, 2, & 3	1,353	721.27 (420.09)	-309W to 2000W

The following variable was appended to the spreadsheet of essay scores provided by ACARA on June 25, 2013.

- Column AA – LWA. The value from the Lexile Writing Analyzer.

Predicted Essay Rubric Scores. The natural language processing (NLP) variables that are used by MetaMetrics in the auto-essay scoring (AES) engine can be grouped into four clusters as described below:

- *Content*: Given a technique that reduces a "bag-of-words" into a vector of finite dimensions that roughly describes the meaning of the content, variables are created that describe the length of such vectors (a measure of content knowledge in a chunk of text) and also the angles between them (how related are the chunks of text). Some variables are based on the entire essay and others are based on moving windows of text through the essay to provide some perspective into the flow, style, and cohesion of an essay. Paragraphs are also used as "windows" of text so that a measure of the variance in content between paragraphs can be estimated.
- *Structure*: Frequencies of common structure words and common structure word trigrams (including punctuation) are included to describe the proper use of grammar and punctuation.
- *Entropy*: Measures of entropy for a "bag-of-words" (how rare are the words used) at the essay level and also the mutual information of the "bag-of-words" (how much do these words make sense being used together) are used to estimate information content, the level of vocabulary, and extent to which the vocabulary is related. These variables come from the field of information theory.

- *Surface features*: These variables include the number of words in the essay that exist in a corpus of known words; mean and standard deviation of paragraph length; and the number of words, characters, sentences, and paragraphs. These variables might add only a small amount of predictive power, but are simple to measure. They also have potential to help determine how atypical an essay is and whether it should be flagged to be reviewed by a human.

The variables that are generated by these four clusters are used with a non-deterministic, decision tree-based, machine learning algorithm known as random forest (Breiman, 2001; Segal, 2004; and Strobl, Malley, and Tutz, 2009). Random forest is a well-established machine learning technique that is widely used and considered to be state-of-the-art. Briefly stated, the random forest machine learning algorithm creates ensembles of decision trees where each tree and each node is generated by a random sample of data points and variables. Nodes are split in a way that maximally distinguishes the two sets that result from the split. The predicted dependent variable is the mean of the votes produced by the ensemble of decision trees.

Since scoring the essays is being treated as a "pure" prediction problem, all of the variables within the four clusters were included in the development of the random forest algorithm to create an ensemble of decision tree models for the NAPLAN Total Score (average) and each of the ten marking criterion scores (averages).

Validation of the random forest algorithm within the MetaMetrics AES engine developed to score the NAPLAN essays was conducted using a "cross-validation" approach. In cross validation, some set of the sample is excluded (or "held out") from use in the development/training phase of the process. The excluded sample is then scored using the developed model to test the model (Schneider, 1997). This process can be improved by dividing the sample into multiple subsets and the "holdout" method is completed multiple times (often called the K-fold cross validation). In this study it was straightforward enough carry this division of the sample into multiple sets to the extreme where the "hold-out" set consisted of one essay (often called Leave-one-out cross validation). In this approach, the score for each essay was determined by a model trained on the entire set of essays excluding the essay being scored. By design, the random forest algorithm produces "leave-one-out" models essentially as a by-product and this approach is common. *Table 4* shows the correlations between the predicted ratings for the Total Score and the ten marking criteria and the averages of the ratings provided by ACARA. *Figures 1* and *2* show the relationship between the predicted ratings and the human ratings.

Table 4. Descriptive statistics for the predicted ratings from the MetaMetrics AES engine and the correlations with human raters (Samples 1 & 2, *N* = 1,014).

Essay Rubric Score	Predicted Mean(SD)	Correlation Between Predicted Rating from MM AES and Average Rating
Audience	3.357 (0.970)	0.90
Text Structure	2.304 (0.764)	0.88
Ideas	3.019 (0.784)	0.89
Persuasive Devices	2.371 (0.751)	0.87
Vocabulary	2.687 (0.760)	0.87
Cohesion	2.310 (0.526)	0.80
Paragraphing	1.432 (0.809)	0.86
Sentence Structure	3.048 (0.855)	0.83
Punctuation	2.591 (0.695)	0.74
Spelling	3.713 (0.951)	0.89
Total Score	26.833 (7.804)	0.92

From an examination of the correlations between machine and human raters (range of correlations – from 0.63/0.68 [Rater 1/Rater 2 for Audience] to 0.83/0.82 [Rater 1/Rater 2 for Punctuation]) compared with the correlations from two raters (range of correlations described on page 8 -- from 0.68 [Cohesion] to 0.85 [Spelling]), it can be concluded that the machine (MetaMetrics AES engine) works as well as a human when scoring the ten marking criteria on the NAPLAN essays.

The following variables were appended to the spreadsheet of essay scores provided by ACARA on June 25, 2013.

- Columns AB through AL – Predicted Measures. These are the predicted values from the auto-essay scoring algorithms that were developed for each of the writing scoring criteria (Audience, Text Structure, Ideas, Persuasive Devices, Vocabulary, Cohesion, Paragraphing, Sentence Structure, Punctuation, Spelling), and for the Total score.

Figure 1. Machine versus human scoring of Total Score.

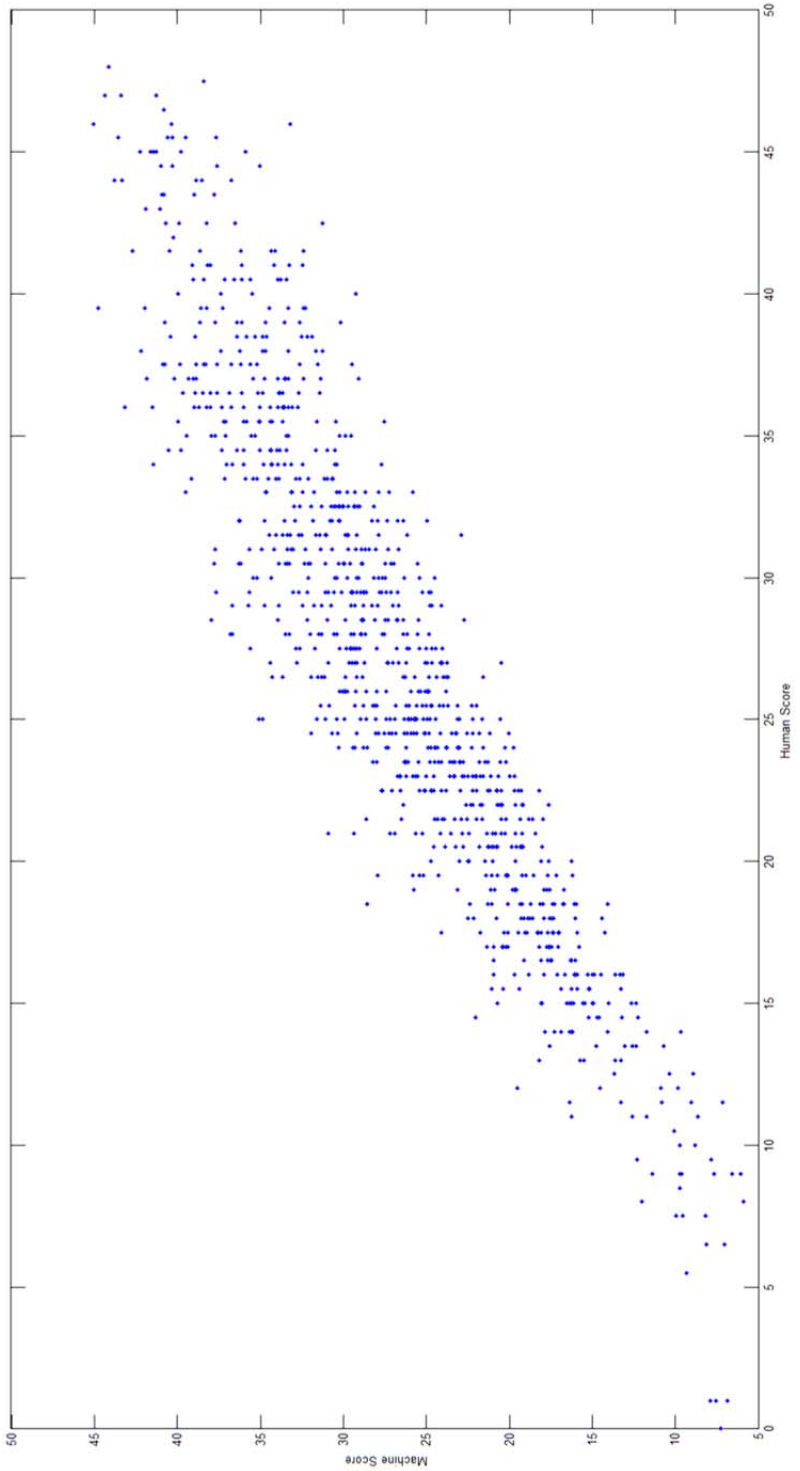
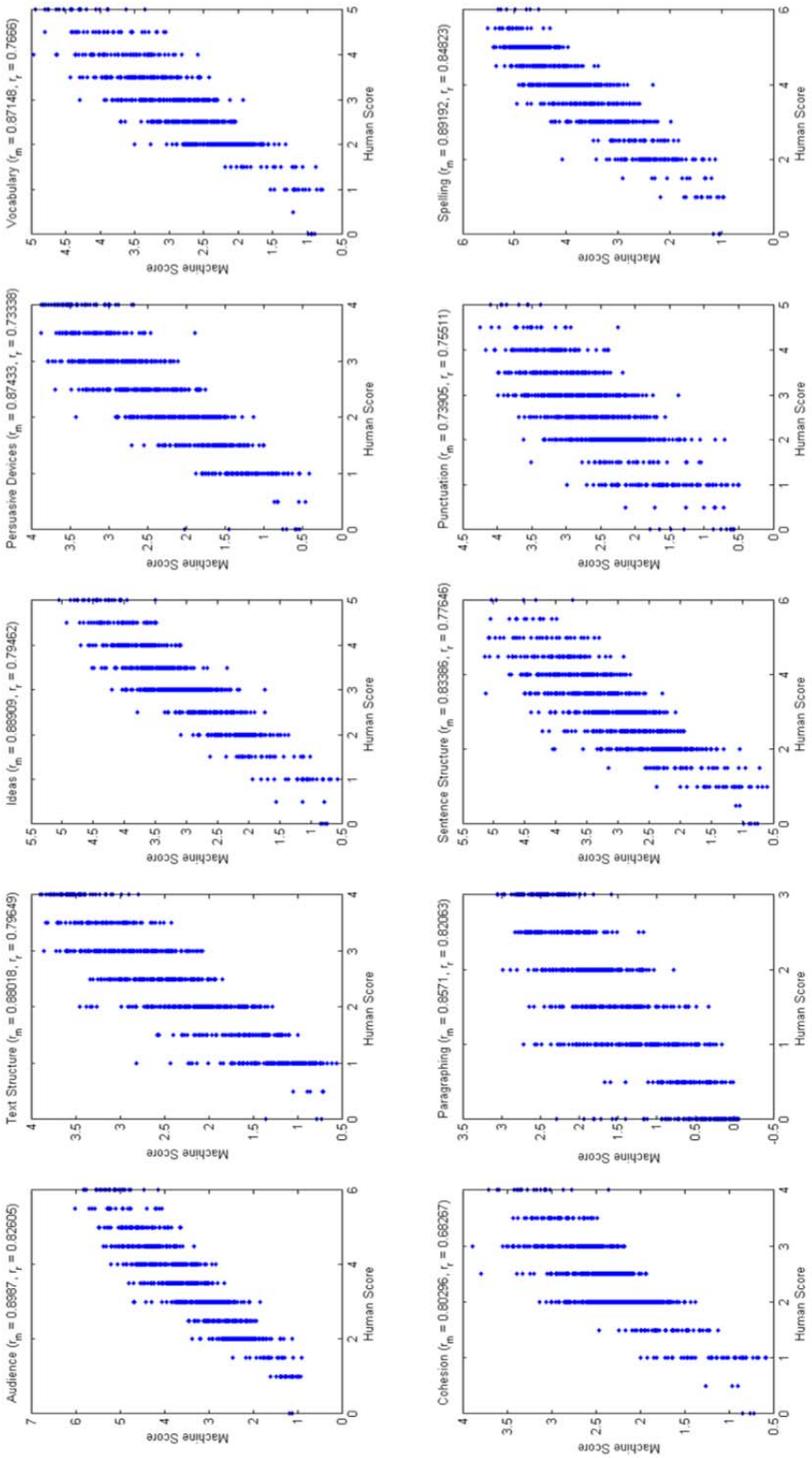


Figure 2. Machine versus human scoring of ten marking criteria scores.



.

Predicted Essay Rubric Ratings. The next step in the process was to determine predicted ratings by developing "cut scores" associated with the rubric scores (integer values only). The "training" involved developing a mechanism to convert the Lexile essay scores, Total predicted scores, and criterion predicted scores into ordinal scales (similar to changing actual temperature measurements into categories of "very hot," "hot," "cool," and "cold"). This was undertaken by developing an iterative procedure that consisted of the following steps:

1. Establish an initial set of cut points that separate the range of scores into categories (NAPLAN hand-scored marks from the rubric).
2. Cycle through the scores associated with the lowest category to find the one that maximizes the Quadratic Weighted Kappa.
3. Retain that measure as the new cut point.
4. Repeat steps 2 and 3 for the remaining categories.
5. Repeat steps 1 through 4 until Kappa delta is less than 0.0001.

For the Total predicted score and the ten criterion scores, two sets of scores for each student (one for each rater) were included in this analysis. All Total scores less than 8 were given an 8 for further analyses (based on a conversation with ACARA staff, June 21, 2013). The following variables were appended to the spreadsheet of essay scores provided by ACARA on June 25, 2013.

- Columns AM through AX – Predicted Ratings. These are the predicted ratings (1) derived from the LWA, the predicted writing scoring criteria, and the predicted Total score and (2) based on the rubrics used by ACARA to score Audience, Text Structure, Ideas, Persuasive Devices, Vocabulary, Cohesion, Paragraphing, Sentence Structure, Punctuation, Spelling, and compute the Total score.

Evaluation of Prediction Models. To evaluate how well the Lexile Writing Analyzer and the predictions from the MetaMetrics AES engine performed, quadratic weighted kappa statistics (Cohen, 1968; Attali, Bridgeman, and Trapani, 2010; Zechner , Higgins, Xi, and Williamson, 2009) for the Sample 1 & 2 essays (*N* = 1,014) essays were calculated (see *Table 5*). Use of quadratic weighting penalizes larger derivations more than smaller deviations.

To provide an approximate interpretive context for these statistics, Landis and Koch (1977) have proposed that, for the unweighted Kappa coefficient, values of 0.61 to .80 indicate substantial strength of agreement and above .80 indicates almost perfect agreement.

Table 5. Evaluation of agreement between machine scoring and human scoring (Samples 1 & 2, *N* = 1,014).

Machine Predicted Rating (Columns AM through AX)	Human Rating (Columns C through Y)	Quadratic Weighted Kappa Statistic
LWA Rating	Average Total Score	0.8828
Total Rating	Average Total Score	0.9190
Audience Rating	Average Audience	0.8548
Text Structure Rating	Average Text Structure	0.8283
Ideas Rating	Average Ideas	0.8320
Persuasive Devices Rating	Average Persuasive Devices	0.8180
Vocabulary Rating	Average Vocabulary	0.8326
Cohesion Rating	Average Cohesion	0.7411
Paragraphing Rating	Average Paragraphing	0.8094
Sentence Structure Rating	Average Sentence Structure	0.8070
Punctuation Rating	Average Punctuation	0.7146
Spelling Rating	Average Spelling	0.8574

While use of quadratic weighting may result in somewhat higher Kappa values, clearly a very strong agreement exists between the Total score based on the human raters and the Total score based on the MetaMetrics AES engine. In addition, these values far exceed the standard of .70 used by Educational Testing Service as a minimum bar that any automated essay scoring system must exceed (Ramineni, et. al., 2012).

MetaMetrics continues to research essay scoring and make enhancements and modifications in the auto-essay scoring engine (AES) employed in this study. While the quadratic weighted Kappa statistic for the punctuation criteria was acceptable, this area will be added to the list of research needed to support the MetaMetrics AES.

References

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment, 10*(3). Retrieved January 27, 2012 from http://www.jtla.org.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2012). 2012 National Assessment Program Literacy and Numeracy: Persuasive writing marketing guide. Sydney, Australia: Author.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2013-1). "Writing." Retrieved on April 22, 2013 from http://www.nap.edu.au/naplan/about-each-domain/writing/writing.html.

Biancarosa, G., & Snow, C. E. (2004.) *Reading Next—A Vision for Action and Research in Middle and High School Literacy: A Report to Carnegie Corporation of New York*. Washington DC: Alliance for Excellent Education.

Breiman, L. (2001). Statistical modelling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Burdick, H., Swartz, C. W., Stenner, A. J., Fitzgerald, J., Burdick, D., & Hanlon, S. T. (in press). Validity of a computer analytic writing ability scale. *Literacy Research and Instruction*.

Cohen J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213-20.

Graham, S., & Perin, D. (2007).*Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-74.

MetaMetrics, Inc. (2008, March). *Implications of the Lexile Framework for Writing for North Carolina's General Model of Writing Assessment*. Durham, NC: Author.

Ramineni, C., Trapani, C.S., Williamson, D.M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the GRE ® issue and argument prompts*. Princeton, NJ: ETS.

Schneider, J. (1997). "Cross validation." Retrieved on July 22, 2013 from http://www.cs.cmu.edu/~schneide/tut5/node42.html.

Segal, M.R. (2004, April). *Machine learning benchmarks and random forest regression*. Center for Bioinformatics and Molecular Biostatistics, UC San Francisco. Retrieved on August 21, 2012 from http://escholarship.org/uc/item/35x3v9t4.

Shermis, M.D & Hamner, B. (2012, April). *Contrasting state-of-the-art automated scoring of essays: Analysis*. Paper presentation at the annual meeting of the National Council of Measurement in Education, Vancouver.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random Forests. *Psychological Methods, 14,* 323-348.

Swartz, C.W., Burdick, B., Stenner, A.J., Burdick, D.S., Hanlon, S.T., & Hooper, S.S. (2010). *Implications of the Lexile Framework for Writing for the assessment of written expression* (Research Report). Durham, NC: MetaMetrics, Inc.

Zechner , K., Higgins, D., Xi, X., & Williamson, D.M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2012 from https://www.ets.org/research/topics/as_nlp/speech/.