An Analysis of 2000 Essays Automatically Scored for the New Zealand Qualifications Authority (NZQA) by Rembiont Pty Ltd (Rembiont)

Contract Reference: 10116

August 2021

Ewan Thompson & Dr Robert Williams

Executive Summary

This report provides an analysis of the Automated Essay Scoring Trial conducted by Rembiont Pty Ltd. Rembiont uses an algorithmic approach to essay grading which does not require training in contrast to deep learning AI approaches. Rembiont graded 2,000 essay responses provided by NZQA for Level 1 English and History standards. The grading results were then analysed against human scores provided by NZQA. This report details the findings of this analysis together with various supporting charts and statistics.

Analysis shows that agreement between the Rembiont automated scores and the human scores was lower than our initial expectations. Correlation was particularly poor with the Paragraphing and Sentence Structure components of the Rembiont AES score. When omitting these problematic components from the composite AES score agreement was improved considerably, however was still lower than we have achieved with other large sets of essays.

The Kappa measures between the AES predicted scores and human scores show moderate agreement for English subject essays but only fair agreement for History essays. This indicates potential for using Rembiont AEG to improve NZQA's process for English essay scoring, however Rembiont AEG is not ideally suited for use with NZQA History assessments.

Introduction

The NZQA commissioned the Centre for Research in Applied Measurement and Evaluation at the University of Alberta in 2020 to conduct research into the use of automated essay scoring (AES) methods for evaluating the written-response results produced by New Zealand students in English and History as part of the National Certificates of Educational Achievement (NCEA) examination program. NCEA are qualifications earned by senior secondary students which serve as a credential that can be used for employment applications and university admissions.

As a result of favourable outcomes from this research, NZQA invited commercial AES vendors to submit proposals for further trials of the AES technology. Rembiont submitted a successful proposal for a trial and was subsequently awarded a contract to proceed with an AES trial. This trial involved automated blind-scoring of 2,000 essay responses for Level-1 English and History standards.

Contents

Terminology

AEG – Automated Essay Grading – automated analysis and grading of essays using computer technology.

AES – Automated Essay Scoring – automated generation of scores using computer technology.

NER – Named Entity Recognition – an NLP process to locate and classify named entities in unstructured text into pre-defined categories such as person names, organizations and locations.

NLP – Natural Language Processing – computer technology to process and analyse natural language data.

POS – Part of Speech – Tags that identify speech elements such as nouns and verbs.

Tagging – The process of applying POS tags to each word in unstructured text.

Process

NZQA provided Rembiont with 2,000 essays for the trial broken down as follows:

Standard	Question	Number of Essays for Trial
AS90849 – English 1.1	Q3	500
AS90850 – English 1.2	Q1	500
AS90850 – English 1.2	Q3	500
AS91005 - History 1.5	Q1	500

Rembiont technology grades narrative and persuasive essays of any word length on any topic without the need for training essays. The Rembiont AEG engine uses an algorithmic approach rather than relying on deep learning models, which generally need to be trained on a large set of human-marked essays to obtain accurate results. Although our AEG engine uses some deep learning models for specific NLP tasks such as POS tagging and NER, these models are generic to any English language text and have been pre-trained on a large corpus of English language documents.

Our algorithmic approach to essay grading is based on the results of two PhD research projects at an Australian university. It involves analysing events within the text and applying Information Theory to determine how well narratives and arguments are developed in the essay.

For narrative essays, scoring is based on detecting events in an essay. An event consists of an actor, action and state:

- Actor - a character, mentioned either by name or anaphora, as part of the story
- Action - an act that is performed by an actor
- State - the location, time or condition of the respective actor
- Condition - refers to the physical or mental state of the actor

For persuasive essays, scoring is based on Information Theory techniques to determine the information content of the essay. This is done by determining the root concepts used in an essay via a thesaurus, the concept map is subsequently reduced to a numerical vector. Vector algebra is then used to produce an information content metric.

In addition to Events and Information Content, the system uses various quantitative data to calculate essay scores, such as counts and ratios based on the number of words, nouns, verbs, adverbs, adjectives, simple connectives, advanced connectives, spelling errors and grammatical errors.

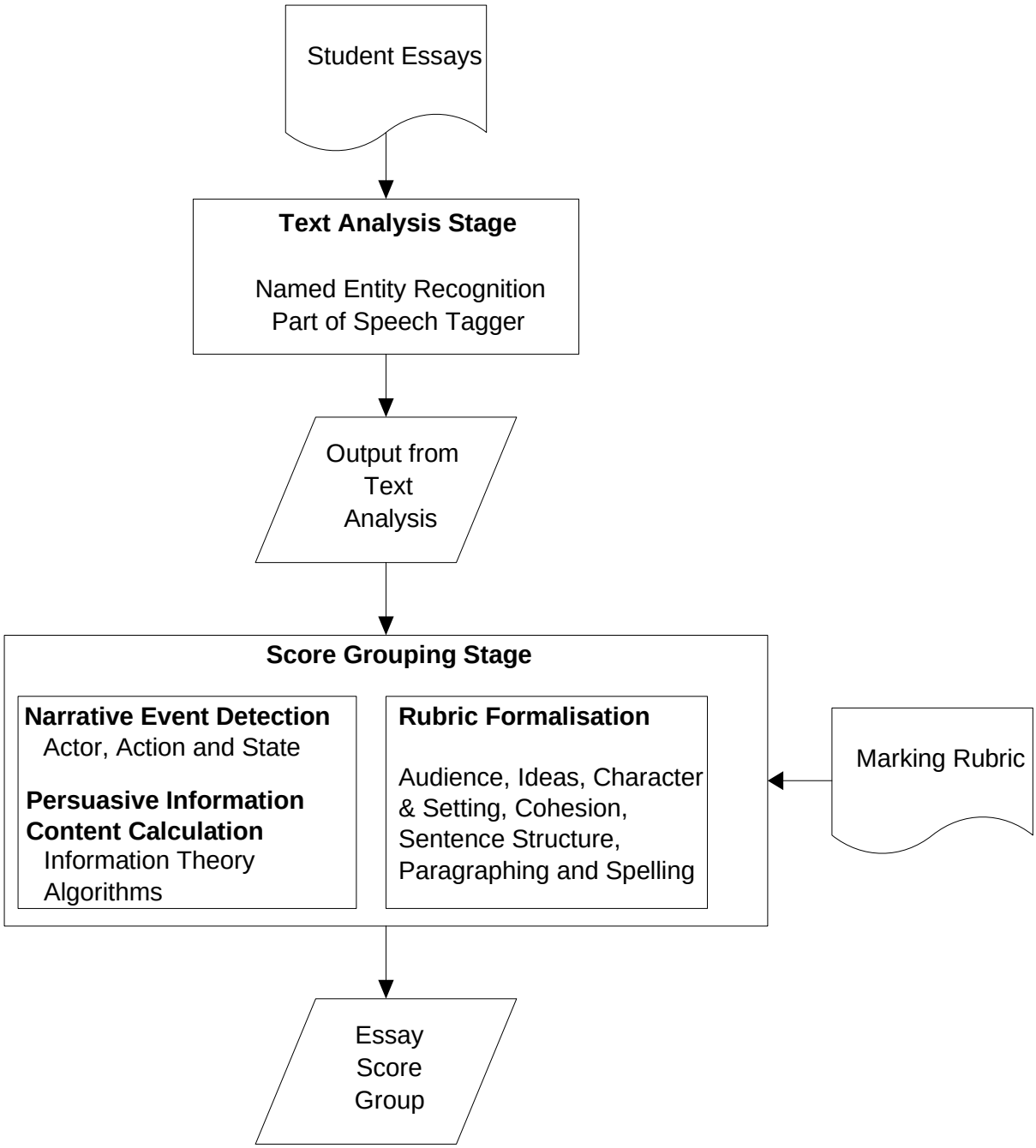
Grading for Content determines the degree to which an essay is on-topic or off-topic. When grading for content, a propriety representation of the knowledge contained in a model answer is generated using thesaurus concept numbers. A model content essay is needed to support this process – the system does not look for key words but thesaurus concepts. Vector Algebra is then used to compute the "closeness" of a student essay to the model content answer.

The system was originally developed to grade essays for the Australian NAPLAN writing assessment. The overall score is a composite of individual scores for various NAPLAN assessment criteria: Audience, Ideas, Cohesion, Character and Setting, Sentence Structure, Paragraphing and Spelling. Our assumption for this trial was that the NAPLAN criteria listed above captured key areas that would indicate the overall quality of an essay and could therefore be used for essays of all types and origins.

Configurable grading rubrics are used to tailor the grading process to customer's requirements. The rubrics specify the type of essay (Narrative or Persuasive), which grading criteria is to be included in the grading process, the score ranges and weightings to be assigned to each criterion, and the cut-score buckets.

When grading an essay, the system undertakes the following tasks:

- Deconstruction of the essay into its constituent parts e.g. paragraphs, sentences, phrases, words, syllables.
- Identification of people, locations and times.
- Tagging parts of speech e.g. nouns, verbs, etc.
- Detection of events
- Determining information content.
- Identifying Spelling and Grammatical Errors
- Production of "internal" scores for Audience, Ideas, Cohesion, Character and Setting, Sentence Structure, Paragraphing and Content.
- Mapping of "internal" scores to output scores based on rubric specifications.



The trial involves the following steps:

- Account and rubric setup (Rembiont) - Completed
- NZQA supply essays to Rembiont (NZQA) - Completed
- First-pass grading run (Rembiont) - Completed
- Prepare interim presentation report (Rembiont) - Completed
- Interim report presentation (Rembiont and NZQA)
- Investigate any issues, anomalies and outliers (Rembiont) - Completed

- Final grading run and provide NZQA with grading results (Rembiont) - Completed
- NZQA provide Rembiont with human scores (NZQA) - Completed
- Prepare final presentation report with human to machine correlation analysis (Rembiont) - Completed
- Final report presentation (Rembiont and NZQA)

Trial Stages 1 – EAS Grading of human-scored exemplars

Rubrics were configured to support the four NZQA cut-score buckets: Not Achieved, Achieved, Achieved with Merit and Achieved with Excellence. Persuasive Grading was specified for all essay batches as this was deemed more suitable than Narrative Grading given the specifics of each question. The "Character and Setting" criterion was disabled as this is only relevant to narratives. NZQA had advised that grading for spelling was not required, this was disabled in the rubric. The resulting grading criteria consisted of five criteria: Audience, Ideas, Cohesion, Sentence Structure and Paragraphing.

NZQA provided sets of exemplars for each Standard/Question, each set contained one human-graded exemplar for each score (0 to 8). Rembiont used these exemplars to benchmark grading results by setting appropriate weightings in the configurable rubrics.

Trial Stage 2 – AES Grading of 2,000 responses for blind scoring

NZQA provided the 2,000 responses to be blind-scored. These were then graded using the weightings established when benchmarking the exemplars in Stage 1.

Agreement Measures

The following measures were used to evaluate the accuracy of the AES score predictions:

- Pearson's Correlation. "A measure of linear correlation between two sets of data" - Wikipedia https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)
- Quadratic Weighted Kappa. A form of Cohen's Kappa which is weighted with a quadratic matrix. Cohen's Kappa is "a statistic that is used to measure inter-rater reliability" – Wikipedia https://en.wikipedia.org/wiki/Cohen%27s_kappa#Weighted_kappa

- Exact Agreement. The ratio of the number of AES predicted scores that are identical to the human score.
- Adjacent Agreement. The ratio of the number of AES predicted scores that are within one point of the human score.
- Cut Score Agreement. The ratio of the number of AES predicted scores that are categorised into the same cut-score group as the human score.

Results

Trial Stage 1 – EAS Grading of human-scored exemplars

An initial grading was performed to obtain metrics to be used in score benchmarking, i.e. aligning Rembiont internal scores with NZQA human scores). This initial grading resulting in good human/AES score correlation for three out of the four batches.

The 90850 Q3 batch had a poor correlation. This was the result of one outlier, the first response in the batch was assigned a human score of zero though our system scored this response much higher. This appeared to be a reasonable essay that would not warrant a zero score, consequently we referred this response to NZQA for investigation. In response, NZQA supplied a substitute 0 score essay and the batch was re-graded, resulting in good correlation.

Correlations and various other measures for the exemplars are shown in the following table. Note that correlations are artificially high as each batch had a small sample size, weightings were established based on the whole batch, and content models were based on the top-scored essays, thus artificially favouring higher scores for these essays.

Measure	90849 Q3	90850 Q1	90850 Q3	91005 Q1
Pearson's Correlation	0.86	0.90	0.93	0.91
Quadratic Kappa	0.84	0.87	0.86	0.83
Exact Agreement	0.44	0.44	0.22	0.33
Adjacent Agreement	0.78	0.78	0.78	0.78
Cut Score Agreement	0.67	0.56	0.44	0.67

Trial Stage 2 – AES Grading of 2,000 responses for blind scoring

An initial AES grading was performed on the 2,000 responses supplied by NZQA. An analysis of the results highlighted a problem with content grading, the scores for Content were irregular and appeared to skew the results. We believe that this is because the essays were not confined to specific topics, for instance it appears that students were allowed to choose from a large variety of different movies for the English essays, and the History essays covered a number of different historical events.

We subsequently re-ran the grading with content checking disabled. This re-run produced better results as indicated by score distributions and mean

scores. The results for all four response batches exhibited score distribution curves that were bell-shaped with minor irregularities. Mean scores were only slightly higher than expectation, except for 91005 Q1 where the mean was significantly higher. There is a clustering of scores evident at the Rembiont internal score of 12.

Apart from the issue described above with content grading, we did not encounter any other major issues whilst processing the essays and all input data was in the expected format.

Stage 3 - Automated score to Human Score Analysis

Following the Stage 2 blind-scoring, Rembiont supplied the AES scores to NZQA, and subsequently NZQA supplied human scores to Rembiont for analysis.

The results of our analysis of the Human vs Rembiont scores are summarised in the following table:

Measure	90849 Q3	90850 Q1	90850 Q3	91005 Q1
Pearson's Correlation	0.57	0.68	0.62	0.53
Quadratic Kappa	0.35	0.45	0.45	0.30
Exact Agreement	0.17	0.15	0.23	0.12
Adjacent Agreement	0.48	0.49	0.57	0.42
Cut Score Agreement	0.31	0.31	0.37	0.30

These measures were lower than our expectations. Our analysis revealed that correlation was particularly poor with the Paragraphing and Sentence Structure components of the Rembiont AES score:

Pearson's Correlation	90849 Q3	90850 Q1	90850 Q3	91005 Q1
Paragraphing	-0.27	-0.12	-0.24	-0.40
Sentence Structure	0.10	0.25	0.14	0.13

When omitting these problematic components from the composite AES score, agreement was improved for all batches over all measures. For convenience in this report, we have named this score subset “C3” as it is comprised of the three categories Audience, Ideas and Cohesion. The result of our analysis of the Human vs Rembiont “C3” scores is summarised in the following table:

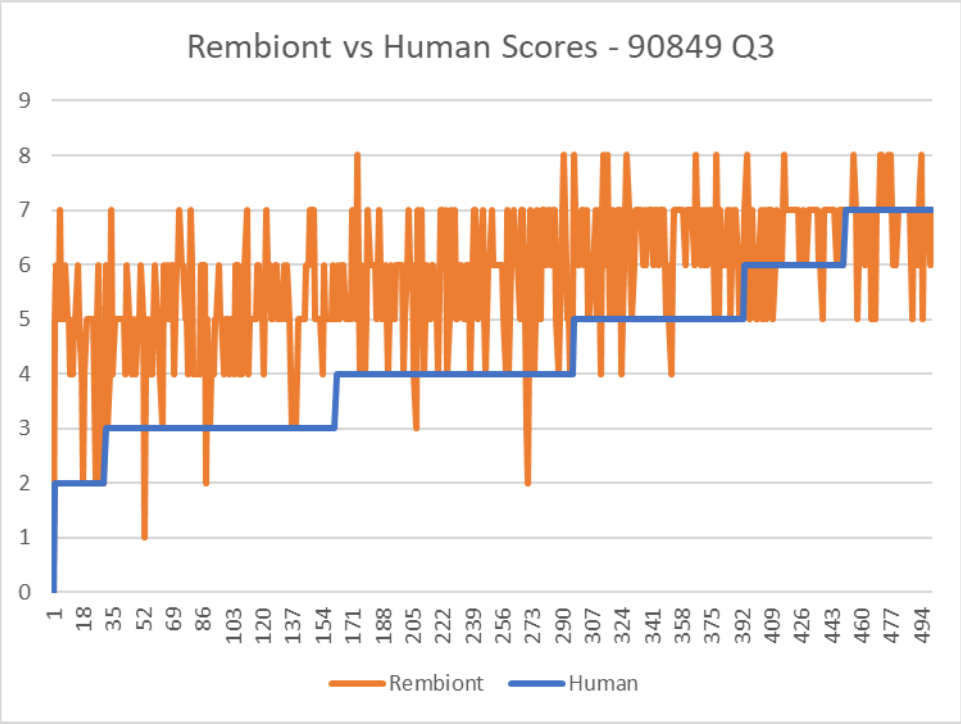
Measure	90849 Q3	90850 Q1	90850 Q3	91005 Q1
Pearson's Correlation	0.61	0.70	0.64	0.61
Quadratic Kappa	0.55	0.57	0.55	0.41
Exact Agreement	0.27	0.23	0.30	0.16
Adjacent Agreement	0.75	0.64	0.67	0.50
Cut Score Agreement	0.55	0.45	0.48	0.34

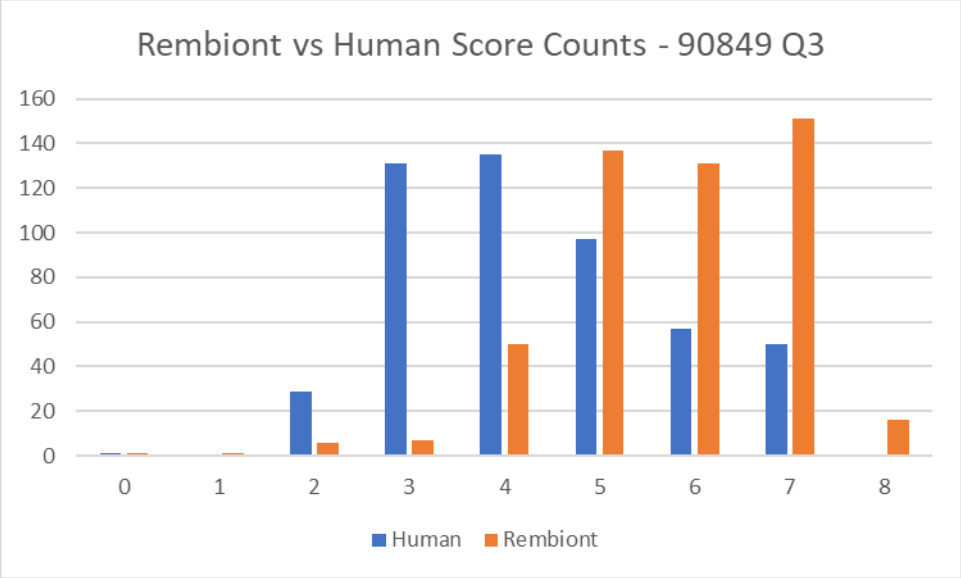
Even with this correction, correlations were still lower than we have achieved with other large sets of essays, typically we see correlations of between 0.80 and 0.90. Potential reasons for the low correlations are discussed later in this report.

AS90849 Q3 Results

For the AS90849 Q3 batch, comparison of Rembiont AES predicted scores against Human scores resulted in the following measures:

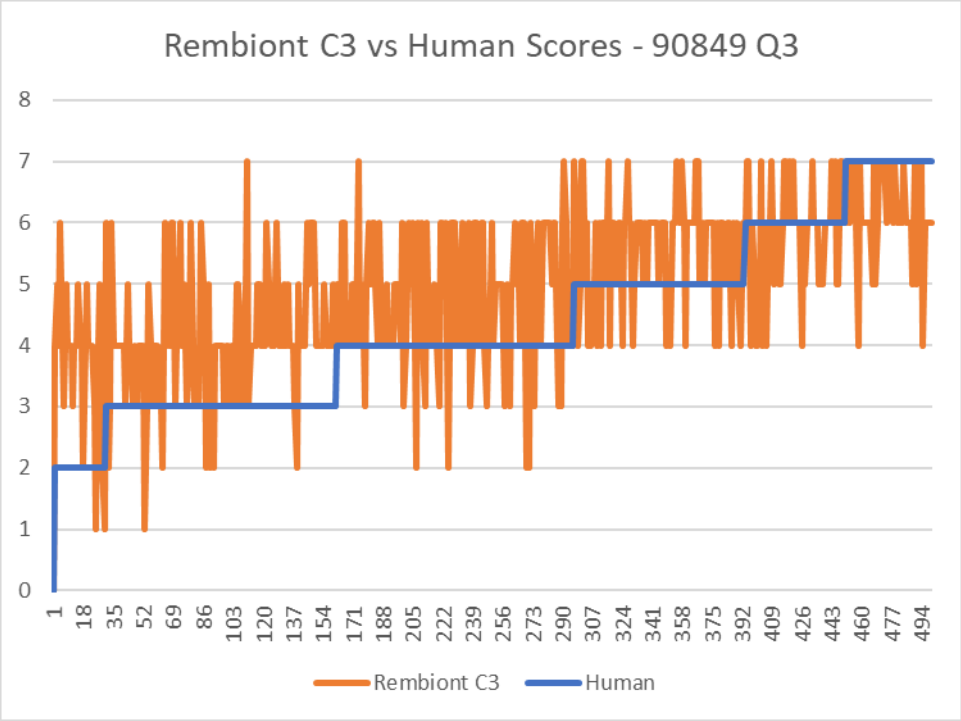
- Pearson's Correlation was 0.57
- Quadratic Weighted Kappa was 0.35 (fair agreement)
- Exact Agreement was 0.17
- Adjacent Agreement was 0.48
- Cut-Score Agreement was 0.31

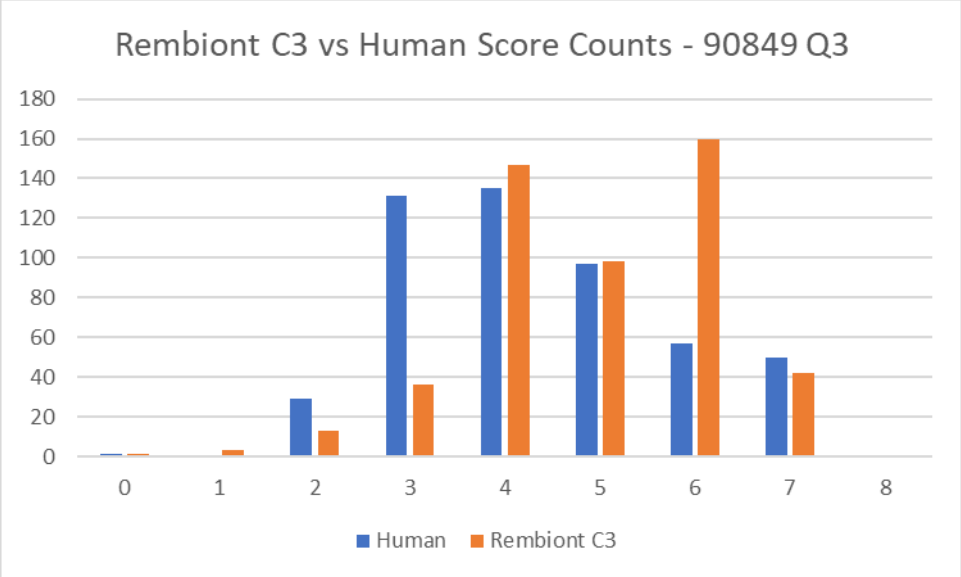




When the problematic Sentence Structure and Paragraphing components were removed from the composite predicted scores, agreement measures improved substantially. Comparison of the resultant Rembiont "C3" predicted scores against Human scores resulted in the following measures:

- Pearson's Correlation was 0.61
- Quadratic Weighted Kappa was 0.55 (moderate agreement)
- Exact Agreement was 0.27
- Adjacent Agreement was 0.75
- Cut-Score Agreement was 0.55



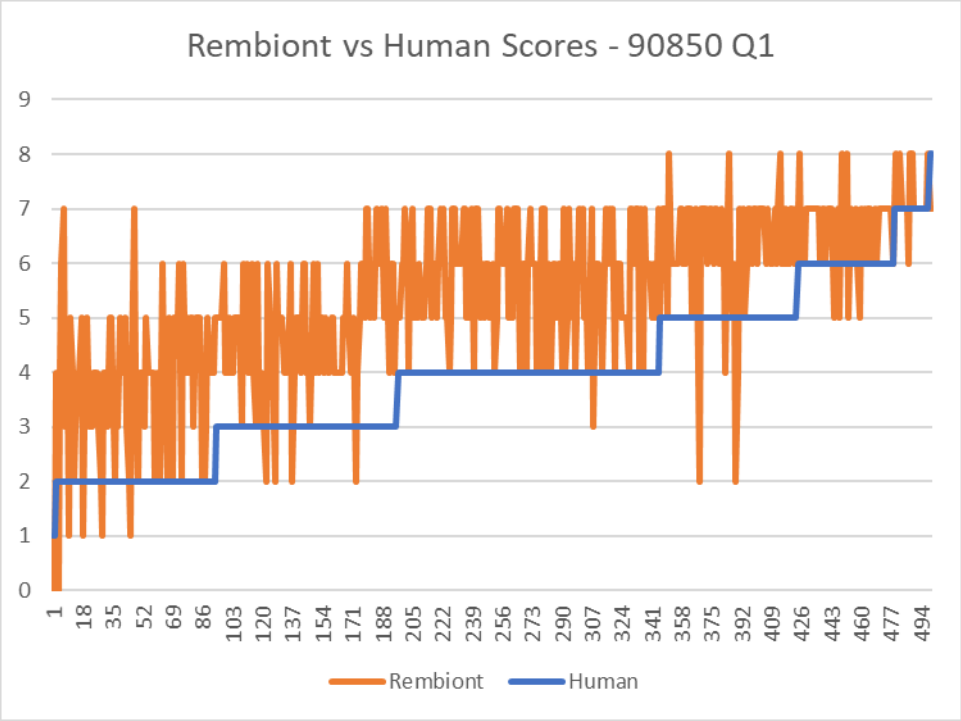


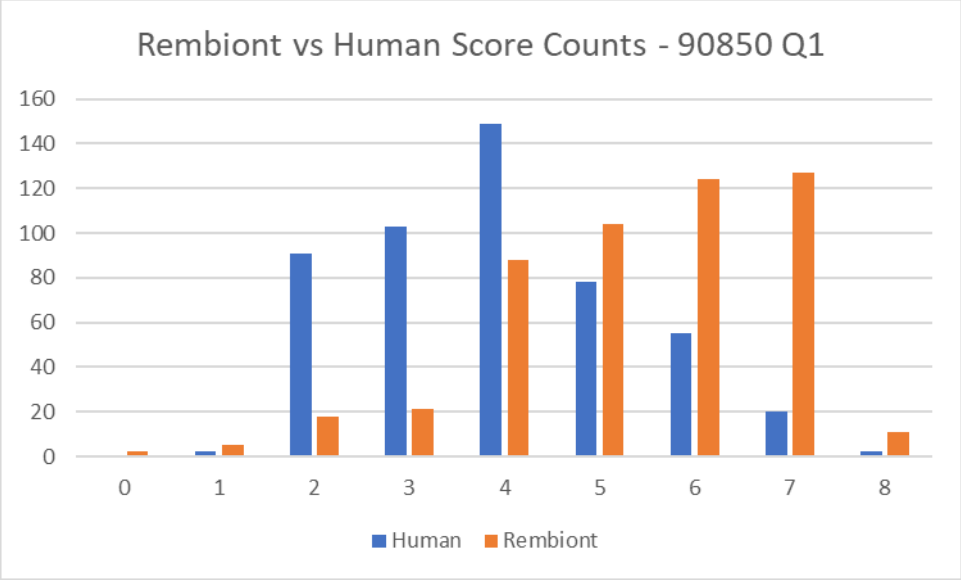
Statistic	*Rembiont*	*Human*	*Rembiont C3*
Mean	5.78	4.336	4.938
Standard Error	0.054746638	0.062722214	0.05687514
Median	6	4	5
Mode	7	4	6
Standard Deviation	1.224172044	1.402511346	1.271766795
Sample Variance	1.498597194	1.967038076	1.617390782
Kurtosis	1.097739918	-0.601431715	-0.01825689
Skewness	-0.738261804	0.341053455	-0.44679659
Range	8	7	7
Minimum	0	0	0
Maximum	8	7	7
Count	500	500	500

AS90850 Q1 Results

For the AS90850 Q1 batch, comparison of Rembiont AES predicted scores against Human scores resulted in the following measures:

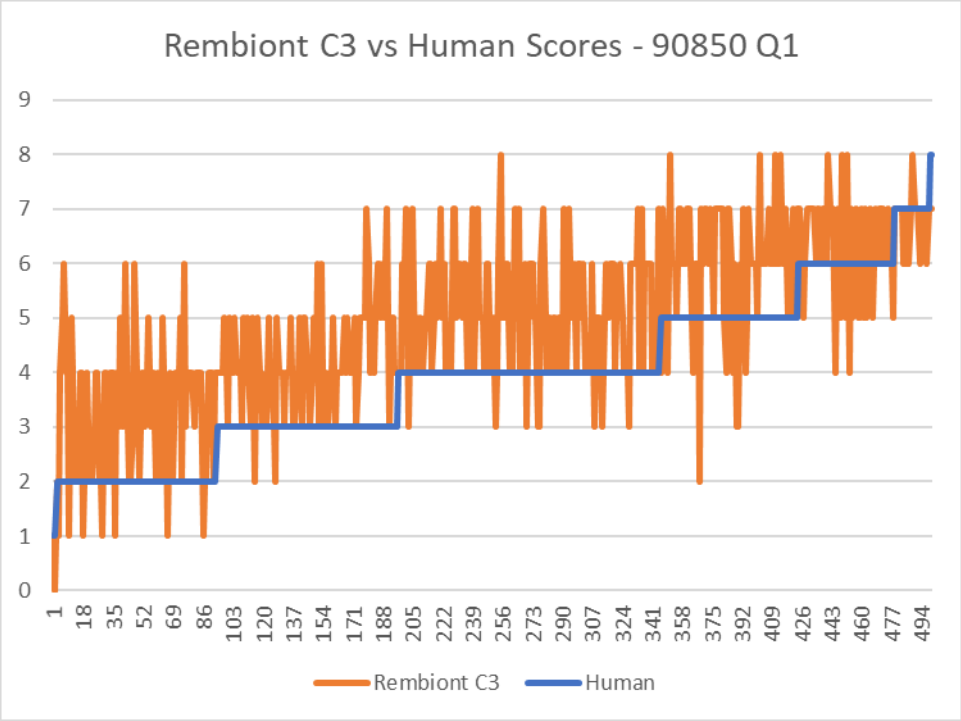
- Pearson's Correlation was 0.68
- Quadratic Weighted Kappa was 0.45 (moderate agreement)
- Exact Agreement was 0.15
- Adjacent Agreement was 0.49
- Cut-Score Agreement was 0.31

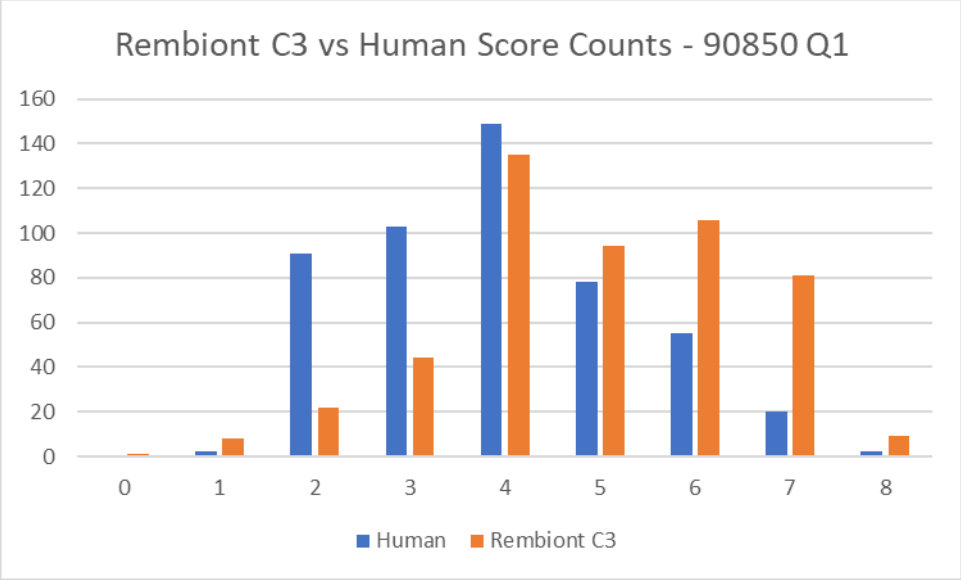




When the problematic Sentence Structure and Paragraphing components were removed from the composite predicted scores, agreement measures improved substantially. Comparison of the resultant Rembiont "C3" predicted scores against Human scores resulted in the following measures:

- Pearson's Correlation was 0.70
- Quadratic Weighted Kappa was 0.57 (moderate agreement)
- Exact Agreement was 0.23
- Adjacent Agreement was 0.64
- Cut-Score Agreement was 0.45



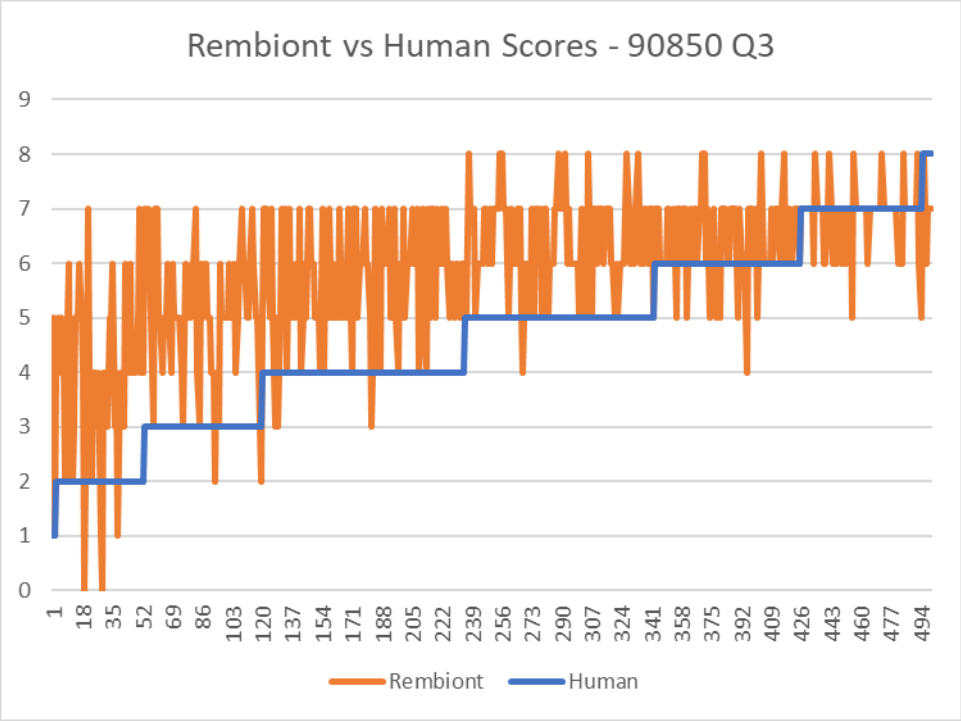


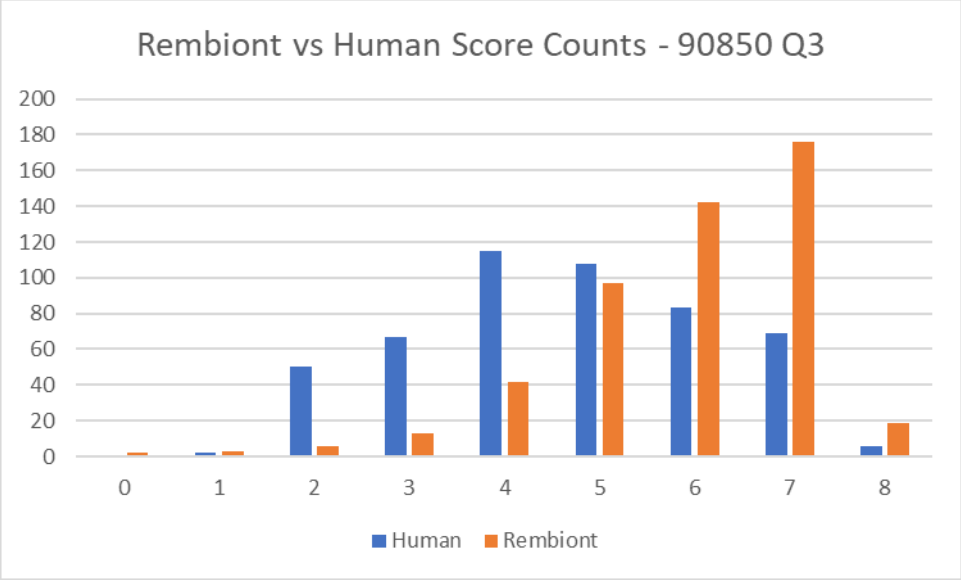
Statistic	*Rembiont*	*Human*	*Rembiont C3*
Mean	5.394	3.93	4.938
Standard Error	0.066921484	0.063072627	0.068510579
Median	6	4	5
Mode	7	4	4
Standard Deviation	1.496409865	1.410346821	1.531943112
Sample Variance	2.239242485	1.989078156	2.346849699
Kurtosis	0.354967426	-0.493596023	-0.339705842
Skewness	-0.736367689	0.365508708	-0.28153492
Range	8	7	8
Minimum	0	1	0
Maximum	8	8	8
Count	500	500	500

AS90850 Q3 Results

For the AS90850 Q3 batch, comparison of Rembiont AES predicted scores against Human scores resulted in the following measures:

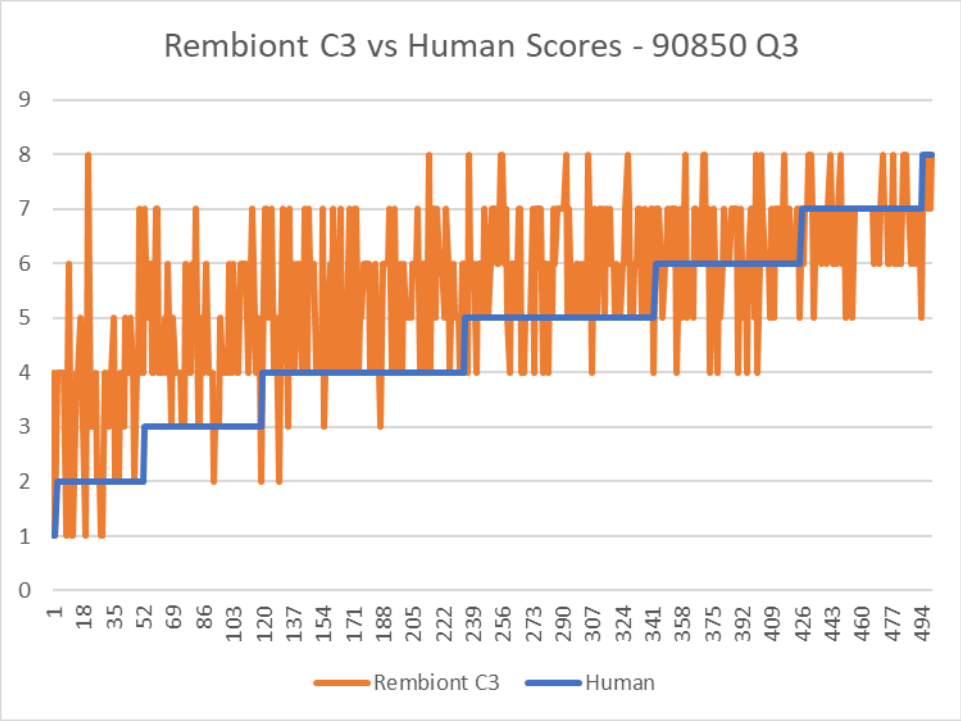
- Pearson's Correlation was 0.62
- Quadratic Weighted Kappa was 0.45 (moderate agreement)
- Exact Agreement was 0.23
- Adjacent Agreement was 0.57
- Cut-Score Agreement was 0.37

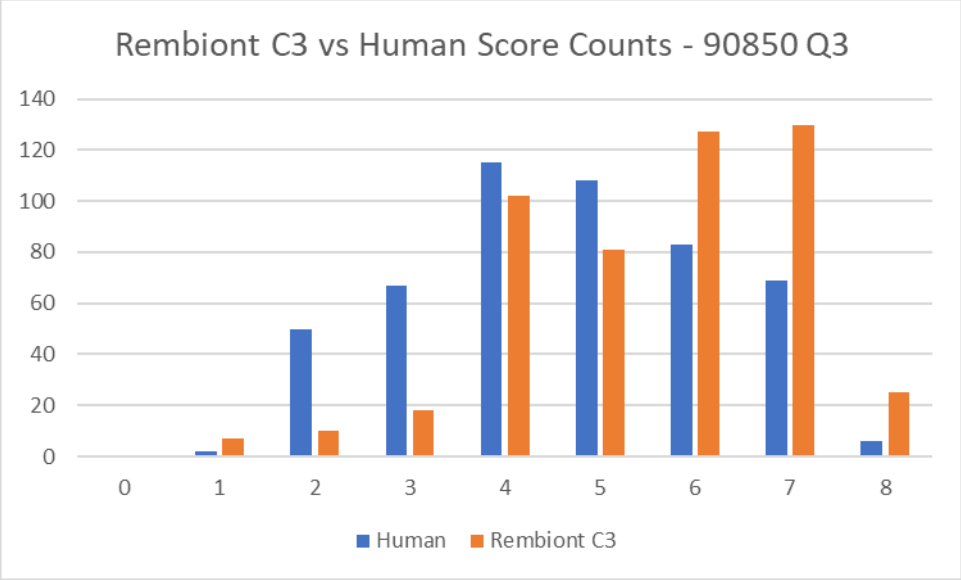




When the problematic Sentence Structure and Paragraphing components were removed from the composite predicted scores, agreement measures improved substantially. Comparison of the resultant Rembiont "C3" predicted scores against Human scores resulted in the following measures:

- Pearson's Correlation was 0.64
- Quadratic Weighted Kappa was 0.55 (moderate agreement)
- Exact Agreement was 0.30
- Adjacent Agreement was 0.67
- Cut-Score Agreement was 0.48



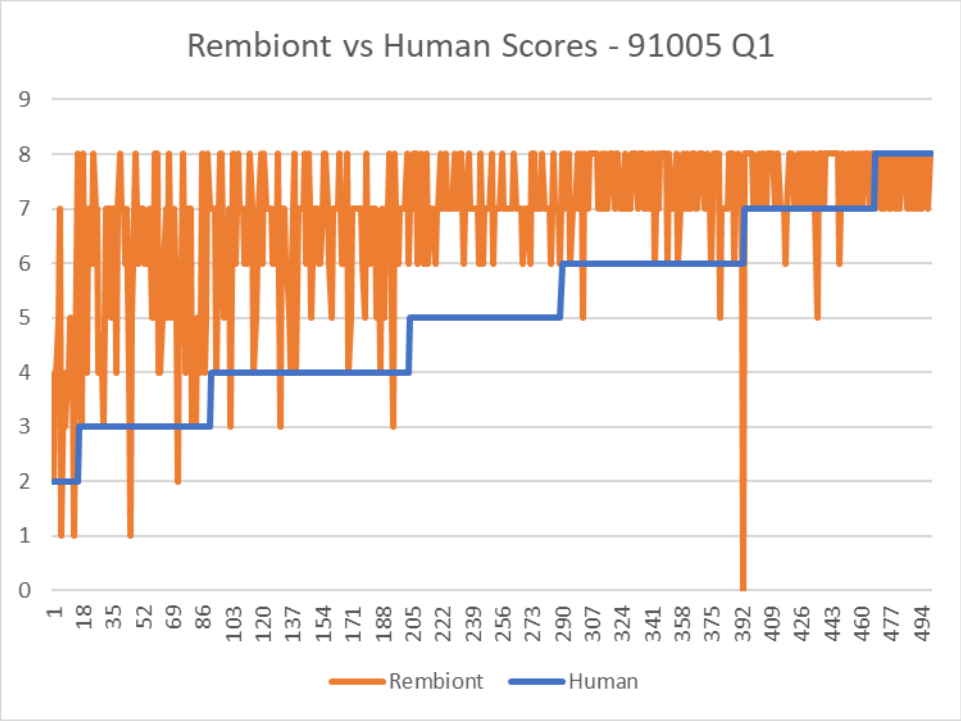


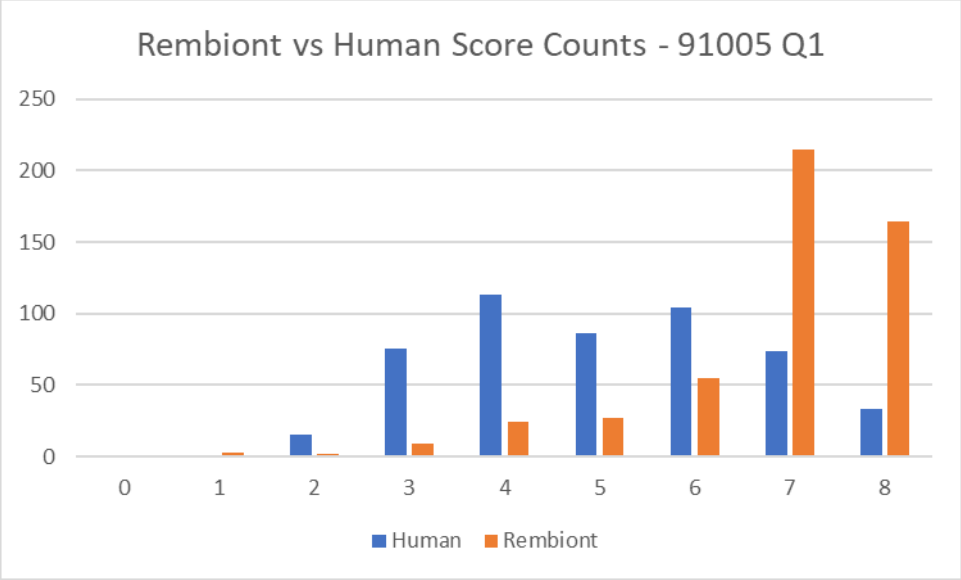
Statistic	*Rembiont*	*Human*	*Rembiont C3*
Mean	5.886	4.664	5.532
Standard Error	0.059067639	0.069971423	0.066594457
Median	6	5	6
Mode	7	4	7
Standard Deviation	1.32079256	1.564608577	1.489097318
Sample Variance	1.744492986	2.448	2.217410822
Kurtosis	2.06180794	-0.825499002	0.015748034
Skewness	-1.177612852	-0.042923308	-0.586955308
Range	8	7	7
Minimum	0	1	1
Maximum	8	8	8
Count	500	500	500

AS91005 Q1 Results

For the AS91005 Q1 batch, comparison of Rembiont AES predicted scores against Human scores resulted in the following measures:

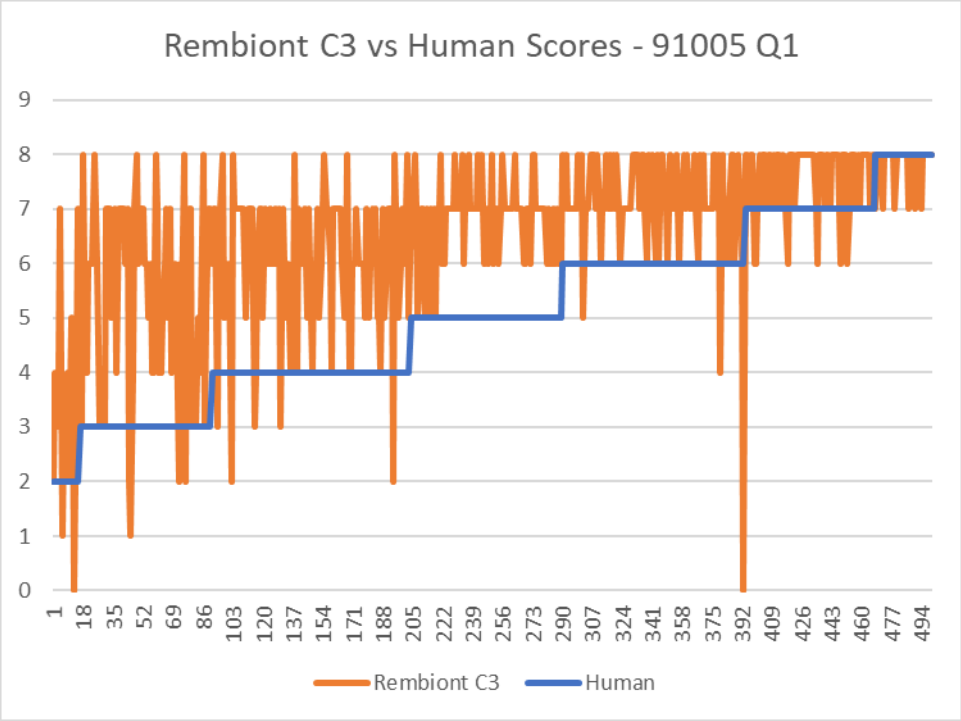
- Pearson's Correlation was 0.53
- Quadratic Weighted Kappa was 0.30 (fair agreement)
- Exact Agreement was 0.12
- Adjacent Agreement was 0.42
- Cut-Score Agreement was 0.30

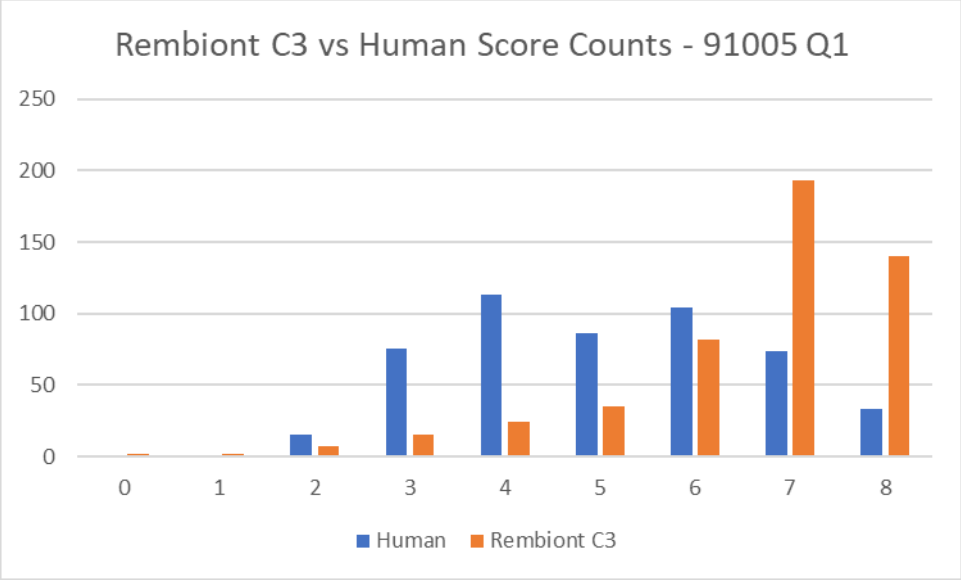




When the problematic Sentence Structure and Paragraphing components were removed from the composite predicted scores, agreement measures improved substantially. Comparison of the resultant Rembiont "C3" predicted scores against Human scores resulted in the following measures:

- Pearson's Correlation was 0.61
- Quadratic Weighted Kappa was 0.41 (moderate agreement)
- Exact Agreement was 0.16
- Adjacent Agreement was 0.50
- Cut-Score Agreement was 0.34





Statistic	*Rembiont*	*Human*	*Rembiont C3*
Mean	6.824	5.086	6.59
Standard Error	0.059338775	0.070534802	0.065760542
Median	7	5	7
Mode	7	4	7
Standard Deviation	1.326855339	1.577206124	1.470450424
Sample Variance	1.76054509	2.487579158	2.162224449
Kurtosis	4.043919849	-0.92602811	2.706658803
Skewness	-1.813460436	0.076117567	-1.557674181
Range	8	6	8
Minimum	0	2	0
Maximum	8	8	8
Count	500	500	500

Analysis Findings

The AES predicted scores were higher than human scores across the board. This indicates that the benchmarking process using the exemplars did not accurately reflect the data ranges of the live batch responses. Analysis of the two data sets shows that word counts and information content were higher in the live batch responses than the exemplars:

Question	Metric	Exemplars	Live
90849 Q3	Max Total Info	886	1277
90850 Q1	Max Total Info	1290	1469
90850 Q3	Max Total Info	880	1498
91005 Q1	Max Total Info	1299	1807
90849 Q3	Avg Word Count	771	852
90850 Q1	Avg Word Count	690	775
90850 Q3	Avg Word Count	747	903
91005 Q1	Avg Word Count	1018	1216
90849 Q3	Max Word Count	1403	1999
90850 Q1	Max Word Count	1765	2611
90850 Q3	Max Word Count	1519	2530
91005 Q1	Max Word Count	2233	3328

Revising the benchmark metrics to reflect this difference would better align the predicted and human scores and result in improved Exact Agreement, Adjacent Agreement and Cut Score Agreement, however it would have little impact on Pearson's Correlation and Quadratic Weighted Kappa measures.

When reviewing outliers between the AES and human scores, we were not able to identify clear reasons for the score differences. In some cases the human scores seemed suspect to our untrained eyes. A particularly extreme example is the 91005 Q1 essay with ID of 4925827600. This essay was awarded a human score of 6 however the Rembiont AEG assigned a score of 0. This was a very short and simple essay, and the text appeared to be limited to the essay introduction. The essay was very simplistic compared to other essays with human scores of 6. Perhaps the essay, although being short, captured the essence of the correct answer expected, though this seems unlikely. More likely there was a human error in the scoring process, or else the full essay text was inadvertently truncated during the encoding and data cleansing process. If the later, there may be other examples where AES scores will be incorrect as the AEG was not considering the full essay text.

Some of the outliers may be explained by the fact the Rembiont system is currently unable to detect certain situations that would be obvious to a highly trained human marker, for example:

- Rote-learned essays loosely adapted to the question.
- Inaccurate or made-up quotations from subject texts.
- Plagiarism.
- For history essays, inaccurate facts, times, or sequence of historical events.
- Advanced rhetorical devices that add considerable value and impact to the response.
- Responses that use very basic English language constructs, however succinctly and accurately provide a correct answer and illustrate the student's understanding of the subject.

To assist with validating the results, Rembiont also graded the essays using an alternate Blue Wren system (under permission from Blue Wren Pty Ltd) to identify any outliers between the two systems. The Blue Wren system requires training data, so we used the first 200 essays from each batch for training, and then graded the remaining 300 essays. The scores from the Blue Wren system have slightly better correlation with the human scores, but are still less than results typically obtained from the system. We have produced a report detailing the results of the Blue Wren grading which Rembiont can make available to NZQA at no cost on request. Note that this processing was performed by Rembiont and the NZQA essay data was not provided to any third party.

Conclusion

Our assumption prior to this trial was that scoring criteria and weighting derived from NAPLAN assessments would map well to the NZQA assessments. The low agreement measures resulting from this trial has proved this assumption to be questionable. In particular, our Sentence Structure and Paragraphing grading criteria had low correlation with the NZQA human scores.

Using a revised criteria set, the Quadratic Kappa measures indicate "moderate agreement" between AES and Human scores for the English standards (90849 and 90850). Combined with refined benchmarking metrics, the Rembiont AEG system has potential for use in validating and identifying outliers in NZQA human-scored essays for English assessments. The predicted scores are certainly accurate enough to support an essay writing practice system for New Zealand English students, along the lines of Rembiont's NAPLearn product developed for Australian English students.

Even with a revised criteria set, the Quadratic Kappa measures indicate only "fair agreement" between AES and Human scores for the History standard (91005). The Rembiont AEG system currently does not predict scores with sufficient accuracy to validate or identify outliers in NZQA human-scored essays for History assessments. Better results could be obtained for these essays using AEG systems that utilise a training stage to identify the best features to use as independent variables for prediction.

Automated Essay Grading systems are unlikely to match the insight and accuracy of a highly trained human marker even when using the most advanced AI and Deep Learning techniques available today. Nevertheless, there is considerable scope to use automated scoring to vet large batches of human scores for quality control purposes.

As a final point, it is important to ensure the veracity of the human scores in order to provide a reliable benchmark for AES accuracy. To this end, it is advisable that human scores used for AES verification or training be based on essays that have each been marked by at least two human graders. This may well have been done for this trial, though at the time of writing we were unable to confirm this with NZQA. If the human scores were not double-marked, provision of double-marked scores would in high probability result in improved AES to human score agreement.