Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

18 November 2025

[redacted]

Tēnā koe [redacted]

Official Information Act Request

Thank you for your request of 20 October 2025, under the Official Information Act 1982 (OIA), for the following information:

May 2025 Automated Text Scoring results, particularly the percentage agreement with human markers

Attached, as Appendix 1, is the information relevant to your request. We have provided a summary for ease of reading.

Given your earlier OIA request, we have also included, in Appendix 1 and below, relevant contextual information from the AI marking Pilot which was undertaken using student responses from the September 2024 assessment event (AE2 2024).

Information to supplement Appendix 1

- In the 2024 pilot, 100% of the 35,000 responses were marked by both machine marking (Automated Text Scoring (ATS)) and human markers, producing an 80% agreement rate on Achieved/Not Achieved decisions (as reported to the Minister of Education).

- ATS speeds up result turnaround, giving schools more time to support students who did not achieve between the first assessment event (AE1 2025) and the second assessment event (AE2 2025). For AE1 2025, human marking focused on responses near the achievement boundary (about 36% of assessments) to reduce marking time.

- Because AE1 2025 only had human marking for responses near the achievement boundary (about 36% of assessments), there is no full human-marked dataset. This means we cannot produce an equivalent agreement measure to the pilot across AE1 dataset.

- NZQA used more specific statistical methods in the pilot and AE1 (Pearson correlation coefficient, Quadratic Weighted Kappa, Cohen's Kappa, Matthews Correlation Coefficient, noted in the results table in Appendix 1) to compare ATS and human scores that are not affected by the condensed sample around the achievement boundary to ensure confidence in the marking process and the accuracy of student achievement results. The results tables are included in Appendix 1.

Our response to your request may be published on our website after five working days. Your name and contact details will be removed before publication.

If you require further assistance or believe we have misinterpreted your request, please contact Elizabeth Templeton in the Office of the Chief Executive, email elizabeth.templeton@nzqa.govt.nz or telephone (04) 463 3339.

You have the right to seek an investigation or review by the Ombudsman of this decision under section 28(3) of the Official Information Act 1982. Details of how to make a complaint can be found at www.ombudsman.parliament.nz. You can also telephone 0800 802 502 or write to the Ombudsman at PO Box 10152, Wellington, 6143.

Nāku nā

Dr Grant Klinkum
Pouwhakahaere/Chief Executive

Appendix 1

Pilot Data from September 2024 Assessment

We compared the overall student achievement rate (achieved/not-achieved) of the machine marking with the actual achievement rate based on human marking. The pilot results showed a variance of only 2.1 percentage points between the human-marked achievement rate of 54.5% and the machine results of 56.6% for the same cohort.

At an individual student achievement level, the NDS (the supplier) model has an Agreement Rate of 79.1%, as shown in Figure 1. The shaded cells in Figure 1 are the students for whom there was a discrepancy between the outcome (Achieved (A) vs Not Achieved (NA)) resulting from human marking and machine marking.

Figure 1: Comparison between student outcomes – human marked vs machine marked

		Human marked		
		NA	A	
Machine marked	NA	11,913 (34.0%)	3,295 (9.5%)	15,208 (43.4%)
	A	4,017 (11.5%)	15,779 (45.1%)	19,796 (56.6%)
		15,930 (45.5%)	19,074 (54.5%)	35,004 (100%)

Analysis shows that a high proportion of these discrepancies are close to the cut score boundary between an Achieved and Not Achieved result.

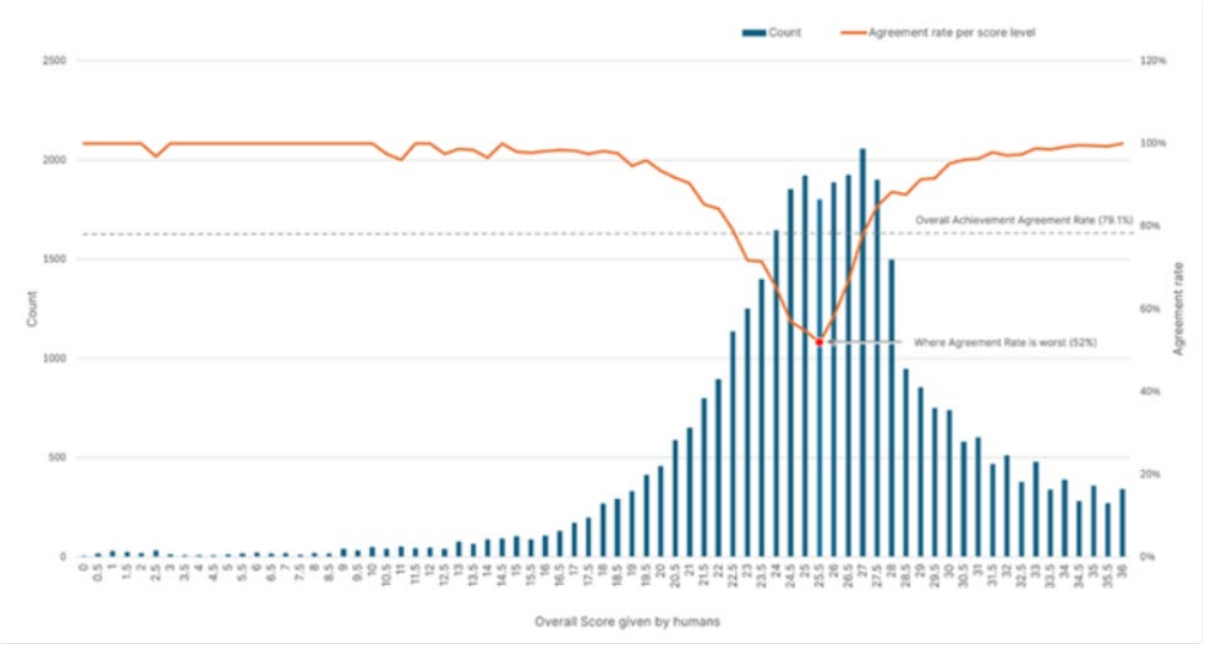
To provide confidence in our automated marking solution, we decided to use a check-marking process. Human marking will be applied when the machine's score is near the achievement boundary.

May 2025 Assessment Event (AE1 2025)

Ahead of AE1 2025, NZQA shared a high-level figure of “80% agreement” between ATS and human marking. This figure was based on pilot data comparing 35,000 student responses and was rounded from an actual agreement rate of 79.1% for ease of communication. The data shows that agreement between ATS and human scores tends to be lower near the achievement boundary (cut score), which makes sense—small differences in scores near this boundary can affect whether a student passes, while the same differences further away don’t have the same impact.

In AE1 2025, human marking was focused around the achievement boundary, so there’s no human-marked data outside that range. Because of this, the 80% agreement rate can’t be recalculated across the full AE1 dataset.

NZQA has used different statistical methods to compare ATS and human scores to ensure confidence in the marking process and the accuracy of student achievement results



A total of 43 experienced human markers were contracted to double mark selected scripts blind (i.e. not being able to see the ATS score). These scripts were identified by targeting the boundary of achievement, as this is the area of most material importance (student achievement outcomes) coupled with a natural dip in the accuracy of ATS scoring (when compared with human scores).

The ATS model also produced a low confidence flag when the confidence in the score the model gave was low. This could be due to reasons such as the response being repetitious or too short for the model to have confidence in the score it gave. Our human contracted workforce marked all of these, irrespective of the boundary.

If only one of a student's pieces of writing met the criteria, both pieces of writing were identified for human marking.

The total number of student responses marked by our human contracted workforce was 20,128 which equates to over 36% of the total number of students who submitted a response for assessment. A small number of additional responses were also human marked to ensure accuracy.

Additionally, several exercises were undertaken to analyse the data and investigate the robustness of the scores.

The following tables summarise the findings for AE1 2025 across a range of statistical measures compared with the pilot. These include for each rubric per question:

- The **Exact** agreement (where rounded marks given by the ATS model and humans are exactly the same)
- The **Adjacent** agreement (where marks given by the ATS model differed from the humans by 1 score point)
- The **Exact + Adjacent** agreement (combining the two to show the total agreement rate within 1 score point)

The average Exact + Adjacent of 0.993 (or 99.3%) was on par with that of the pilot, which had a target of >90%.

However, the Exact agreement of 0.649 (64.9%) is an improvement on the pilot Exact agreement rate of 0.620 (62%).

Figure 1: AE1 2025 statistical findings

AC=Accuracy
CO=Content
LA=Language
ST=Structure

2025 - Assessment Event 1

Item	Sample	Exact	Adjacent	Exact + Adj.	Pearson	Cohen Kappa	Quadratic Weighted Kappa	Matthews Coefficient
W1Q1_AC	10,458	0.565	0.429	0.993	0.500	0.209	0.464	0.229
W1Q1_CO	10,458	0.661	0.328	0.989	0.541	0.170	0.503	0.177
W1Q1_LA	10,458	0.692	0.302	0.995	0.557	0.260	0.545	0.262
W1Q1_ST	10,458	0.675	0.314	0.989	0.579	0.191	0.528	0.199
W1Q2_AC	10,458	0.503	0.49	0.994	0.654	0.182	0.534	0.240
W1Q2_CO	10,458	0.700	0.294	0.994	0.715	0.282	0.678	0.288
W1Q2_LA	10,458	0.728	0.27	0.998	0.735	0.338	0.701	0.351
W1Q2_ST	10,458	0.720	0.276	0.995	0.731	0.326	0.700	0.331
W2Q1_AC	7,982	0.526	0.465	0.991	0.526	0.169	0.442	0.208
W2Q1_CO	7,982	0.729	0.264	0.992	0.609	0.256	0.589	0.263
W2Q1_LA	7,982	0.737	0.258	0.995	0.605	0.291	0.590	0.293
W2Q1_ST	7,982	0.719	0.274	0.993	0.616	0.239	0.587	0.243
W2Q2_AC	7,982	0.474	0.519	0.993	0.638	0.178	0.539	0.214
W2Q2_CO	7,982	0.614	0.376	0.990	0.675	0.231	0.638	0.248
W2Q2_LA	7,982	0.677	0.32	0.996	0.72	0.298	0.694	0.301
W2Q2_ST	7,982	0.659	0.334	0.994	0.705	0.276	0.677	0.285
AVERAGE	**NA**	**0.649**	**0.345**	**0.993**	**0.632**	**0.244**	**0.588**	**0.258**

Figure 2: pilot statistical findings (sample from AE2 2024)

Vantage UniMetric (PILOT Sept'24)

Item	Sample	Exact	Adj.	Exact + Adj.	Pearson	Cohen Kappa	Quadratic Weighted Kappa	Matthews Coefficient
W1Q1_AC	17,505	0.603	0.389	0.992	0.234	0.273	0.435	0.298
W1Q1_CO	17,505	0.600	0.390	0.990	0.213	0.147	0.400	0.153
W1Q1_LA	17,505	0.644	0.351	0.995	0.215	0.219	0.436	0.223
W1Q1_ST	17,505	0.639	0.357	0.996	0.220	0.185	0.437	0.192
W1Q2_AC	17,505	0.603	0.391	0.994	0.208	0.321	0.534	0.357
W1Q2_CO	17,505	0.566	0.424	0.990	0.203	0.182	0.517	0.194
W1Q2_LA	17,505	0.657	0.339	0.996	0.215	0.310	0.575	0.317
W1Q2_ST	17,505	0.612	0.381	0.994	0.208	0.257	0.554	0.266
W2Q1_AC	17,502	0.589	0.401	0.990	0.149	0.272	0.465	0.295
W2Q1_CO	17,502	0.620	0.373	0.993	0.126	0.181	0.447	0.187
W2Q1_LA	17,502	0.640	0.354	0.994	0.121	0.227	0.461	0.231
W2Q1_ST	17,502	0.636	0.359	0.995	0.124	0.196	0.458	0.201
W2Q2_AC	17,502	0.527	0.459	0.986	0.151	0.222	0.473	0.274
W2Q2_CO	17,502	0.647	0.348	0.995	0.148	0.275	0.570	0.286
W2Q2_LA	17,502	0.667	0.327	0.994	0.153	0.320	0.572	0.332
W2Q2_ST	17,502	0.670	0.325	0.995	0.160	0.317	0.586	0.324
AVERAGE	**NA**	**0.620**	**0.373**	**0.993**	**0.178**	**0.244**	**0.495**	**0.258**