Understanding the Relationship between Content Value and Originality

Nick Hope, Innovation Manager, RM Assessment
Mark House, Product Manager, RM Assessment
Hayden Wood, Head of Qualifications and Delivery, ISEB

Contents

Abstract

First presented at the RM Assessment 'AI in Action' Summit in 2025, this case study explores how **RM Assessment** has worked with the **Independent Schools Examinations Board (ISEB)** in exploring new modes of assessment with a focus on the value of original human content in an era dominated by Generative AI. By combining the benefits of Adaptive Comparative Judgement (ACJ) using RM Compare, and content analysis using RM Echo, the study highlights how we might move away from rigid rubrics towards a more holistic means of assessment, whilst maintaining the integrity of students' work now that AI has become so ubiquitous.

Introduction

Since around 2020, the proliferation of commercially available Large Language Models (LLMs) has been dramatic. Suddenly, we are able to prompt a growing number of AI services to:

"Write a story, in the style of a 10 year old, about a chimpanzee who visits New York for the first time."

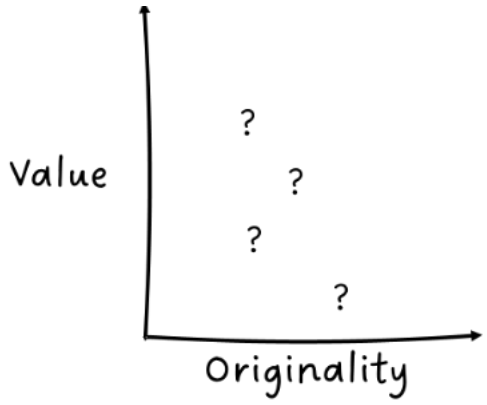
It is rapidly becoming apparent that using written work as a proxy for student's learning might be under serious threat.

The broader context, within which we can explore the implications of 'means of assessment' and 'new technology', is a more fundamental challenge in education and assessment. How educators foster **divergent** versus **convergent** thinking as part of the taught curriculum is critical. Employers increasingly demand more of the former whereby higher levels of creativity and unconventional thinking become the required skills of the modern workforce. On the other hand, evidence points to convergent ways of teaching and assessing where existing methods and rules culminate in a single, correct answer. Generative AI often mimics this same convergent thinking simply because it draws on an existing corpus of information used to train the underlying LLM.

The concern is that if students and teachers become reliant upon AI, that precious idiosyncratic thinking may be even further diminished.

Value versus Originality - with an AI ‘twist’

This study explores the potential relationship between the value that experts place against a piece of student work relative to its originality.



Two hypotheses are evaluated within this context:

> **H1: When judged by human experts, content demonstrating the greatest degree of originality will be valued more highly.**
>
> **H2: Human experts will be able to identify content that has been generated using AI.**

Content is obtained from the ISEB’s annual ‘Time to Write’ competition [1] - a creative exercise giving students of mixed ages complete freedom to produce original stories. Content from 2025 comprises 3,017 stories about the broad theme of ‘Light’.

To address hypothesis **H1**, competition entries are evaluated first using RM Compare [2], an Adaptive Comparative Judgement product allowing two pieces of work at a time to be viewed side-by-side and judged against a holistic statement. This process produces a reliable and stable ‘rank order’ for the work being appraised.

Following the production of a rank order for each of the age categories in the competition, entries are further evaluated using RM Echo, a new content analysis product used to identify areas of similarity between large volumes of written content. [3]

Between the capabilities of these two products, it was anticipated that correlation could be observed according to the study’s hypotheses.

In order to address hypothesis **H2**, 18 AI-generated stories were added to the 3,017 competition entries. The fact that these generated stories were AI-generated was not divulged to the judges, although judges were invited to ‘flag’ any stories that they suspected may have been created using AI.

Findings

The first hypothesis asserted that human judges would rank original content higher than unoriginal content. Results of the similarity analysis from RM Echo show very little similarity across the sample overall. However, items shown to be particularly similar to other items in the sample tend to rank lower according to the results from RM Compare having applied the Adaptive Comparative Judgement (ACJ) algorithm to human judgements

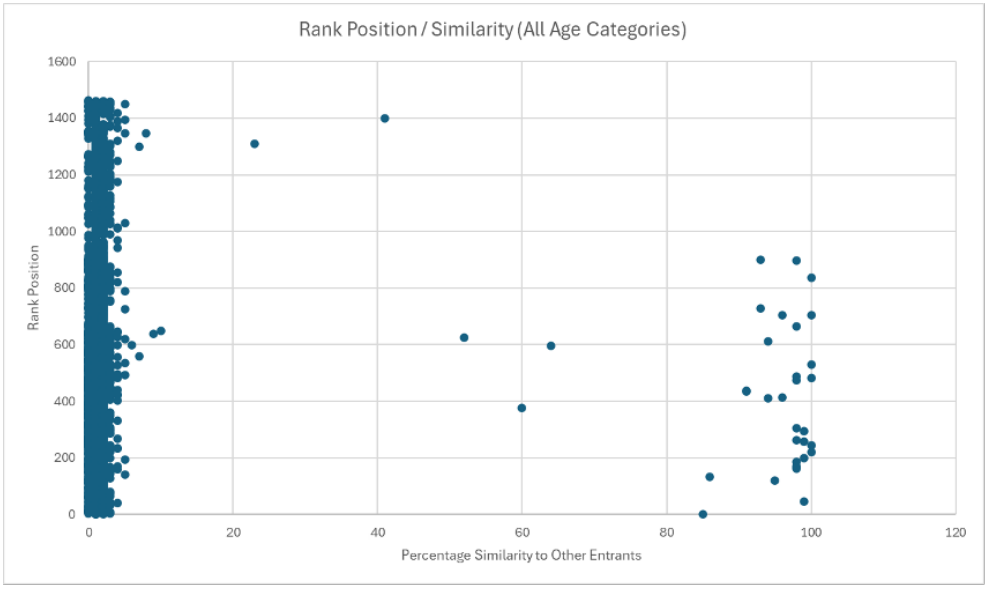


Figure 1: Rank Position against Similarity

Based on a total sample of 3,017 competition entries, data appear to support the hypothesis that original content is valued more highly than content with a high degree of similarity. It should be noted however that some items with a high similarity score may be due to inadvertent duplication of entries where entrants simply submitted the same story twice. This raises questions as to how much we can generalise the results if the majority of items were in fact dissimilar to each other. Furthermore, the nature of the content itself may in fact reduce the likelihood of highly similar items. A creative writing competition is relatively 'low stakes' when compared to summative assessments for formal qualifications and therefore may therefore be less liable to 'malpractice'.

When evaluated by age category, the data show a consistent pattern across entrants whereby most content items appear in Similarity Quartile Group 1 (up to 25% similarity to other items). The data suggest that most content items are reasonably 'original' by this measure, with a small number in Similarity Quartile Group 4 (above 75%) being extremely similar (potentially due to duplication of entries).

		Similarity Quartile			
		1	2	3	4
Rank Quartile	4	160	0	0	3
	3	157	0	0	5
	2	161	0	0	1
	1	161	0	0	1

Table 1: Rank and Similarity Quartiles for Year 3 and 4 age category

		Similarity Quartile			
		1	2	3	4
Rank Quartile	4	364	1	0	0
	3	364	0	0	2
	2	359	0	3	3
	1	360	0	0	7

Table 2: Rank and Similarity Quartiles for Year 5 and 6 age category

Rank Quartile	Similarity Quartile 1	2	3	4
4	224	0	0	2
3	226	0	0	0
2	226	0	0	1
1	221	0	0	5

Table 3: Rank and Similarity Quartiles for Year 7, 8, and 9 age category

Given the low levels of overall similarity between content items, it is difficult to convincingly establish a connection between the value ascribed by human judges and the 'originality' of the items themselves. Although the data are suggestive of a 'high value, low similarity' correlation, a more diverse range of content items demonstrating greater degrees of similarity is required to better establish this relationship.

Hypothesis H1 may therefore be considered *partially validated*.

In the evaluation of hypothesis H2 (identifying AI content), data show that 17 out of 18 of the 'planted' content items eluded detection by the human judges. Interestingly, although these items were not identified as being AI-generated, they tended to be appraised less positively in terms of their overall rank order. This is suggestive that the quality or 'value' of the AI-generated content is generally perceived to be lower.

Rank Quartile	AI Profile Suspected AI	Known AI	Suspected AND Known AI
4	0	0	0
3	5	0	0
2	12	1	0
1	18	5	0

Table 4: Rank Quartiles against AI Profiles for Year 3 and 4 age category

Rank Quartile	AI Profile Suspected AI	Known AI	Suspected AND Known AI
4	2	0	0
3	6	0	0
2	5	3	1
1	9	3	0

Table 5: Rank Quartiles against AI Profiles for Year 5 and 6 age category

Table 5 shows a more nuanced profile whereby the number of content items suspected as being AI-generated is slightly more evenly distributed across the rank order, with 9 items being ranked in the top 50%.

		AI Profile		
		Suspected AI	Known AI	Suspected AND Known AI
Rank Quartile	4	5	0	0
	3	4	0	0
	2	12	3	0
	1	14	3	0

Table 6: Rank Quartiles against AI Profiles for Year 7, 8, and 9 age category

We should exercise caution around the generalisability of these results as the risk of false-positives in any form of 'malpractice' identification when relying on human review alone is hard to ignore. To further evaluate the accuracy of human judgement in the identification of AI-generated content, a future analysis of this sub-sample of 92 items must be completed. New capabilities within RM Echo are expected to be available in the coming months that will help in identifying key linguistic markers of AI-generated content and will be utilised to conduct additional analysis.

Hypothesis H2 may therefore be considered *partially validated.*

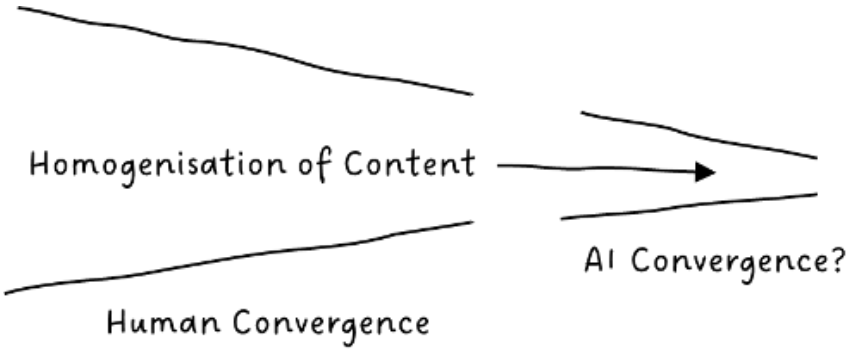
Recognising the limitations of the data, and the lack of functionality within the applied version of RM Echo to further analyse the probability that content is AI-generated, it may be fair to suggest that such content is considered of lower value, but difficult to discern to what extent.

Conclusions

Further research is required to improve our understanding of the connection between content value and content originality, particularly where the use of AI casts doubt on its true source. The limited data within this study demonstrates a cursory relationship between these elements, but greater variety of source material is required. Where this study utilises entries within an existing creative writing competition, more pre-prepared 'planted' content may provide a more representative view of human judgement of value, originality, and use of AI. Equally, applying the same methodology to more 'high-stakes' scenarios may affect results and thus position 'context' as a key variable.

It is noteworthy that, when analysed using RM Echo, content items contained numerous shared examples of phrasing and metaphors. Present across age categories, a number of stories used phrases such as "...voice echoed in my head/mind..." as did one of the 'planted' AI stories. A common thematic conclusion was found in two separate categories: "...testament to the enduring power of the human spirit..." although not in any of the AI stories.

Although a relatively weak correlation, the commonality of phrasing and metaphors across both human and AI-generated content raises some important questions. If the AI-generated content is simply a synthesis of human-created material, then to what extent is the human-created material merely a synthesis of what is taught or simply reflective of how such content is most commonly assessed? Within the context of convergent thinking, it is possible that AI is extending an existing convergence of thinking and writing style.



As LLMs are trained on new content which may itself have been produced with the help of AI, homogenisation of content increases until the degree of originality has been all but eliminated! We should be naturally cautious about generalising these limited results but equally mindful of the potential 'second-order' effects of ever-increasing conformity whether through taught curricula, modes of assessment, or indeed more prevalent use of AI.

This collaborative study between RM and ISEB demonstrates some of the challenges faced when appraising work where 'originality' is highly valued but is increasingly threatened by the capabilities of AI models. It raises important questions relating to how we assess, and the criteria that are considered most important when doing so.

References

1 Time to Write competition

2 RM Compare

3 RM announces new content analysis solution RM Echo

Head office: 142 B, Park Drive, Milton Park, Abingdon, OX14 4SE
Tel: +44 (0)1235 645 316
Web: rm.com/assessment | Mail: digitalassessment@rm.com
Follow us on LinkedIn: RMAssessment