Improving Quality and Safety in AI-Generated Lessons

Hannah-Beth Clark[1,*], Owen Henkel[2], Laura Benton[1], Margaux Dowland[1], Reka Budai[1], Ibrahim Kaan Keskin[1], Emma Searle[1], Matthew Gregory[1], Mark Hodierne[1], William Gayne[1] and John Roberts[1]

[1]Oak National Academy, 1 Scott Place, 2 Hardman Street, Manchester, M3 3AA, United Kingdom
[2]Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom

Abstract

Designing AI tools for use in educational settings presents distinct challenges; the need for accuracy is heightened, safety is imperative and pedagogical rigour is crucial. As a publicly funded body in the UK, Oak National Academy is in a unique position to innovate within this field as we have a comprehensive curriculum of approximately 13,000 open education resources (OER) for all National Curriculum subjects, designed and quality-assured by expert, human teachers. This has provided the corpus of content needed for building a high-quality AI-powered lesson planning tool, Aila, that is free to use and therefore accessible to all teachers across the country. Furthermore, using our evidence-informed curriculum principles, we have codified and exemplified each component of lesson design. To assess the quality of lessons produced by Aila at scale, we have developed an AI-powered auto-evaluation agent, facilitating informed improvements to enhance output quality. Through comparisons between human and auto-evaluations, we have begun to refine this agent further to increase its accuracy, measured by its alignment with an expert human evaluator. In this paper we present this iterative evaluation process through an illustrative case study focused on one quality benchmark - the level of challenge within multiple-choice quizzes. We also explore the contribution that this may make to similar projects and the wider sector.

Keywords

AI-powered lesson planning, Open education resources, LLM as a judge,

1. Introduction

Following the launch of GPT-3.5 in 2022 the edtech market has been flooded with AI tools to support teachers with time-consuming tasks such as lesson planning or generating lesson resources, resulting in a sharp increase in the number of teachers using AI (e.g. Teacher Tapp, 2024). However, there is a lack of evaluation accompanying these tools and the content that they are producing [3]. This is important as these tools are shaping children's education and therefore need to be accurate, safe, context-specific and have research-backed pedagogical design.

Aila is designed to emulate the thought process of an experienced teacher as they plan a lesson. It is intentionally designed not to be a 'single-shot' tool that creates a lesson in one click, but instead supports teacher agency through enabling them to adapt and edit the lesson step-by-step to better suit their students (see Figure 1).

As a publicly funded body in the UK, with the aim of improving pupil outcomes and closing the disadvantage gap, we are in a unique position to innovate within this field. We have created a large corpus of 13,000 Open Education Resources (OER) aligned with the national curriculum for England including; slide decks, worksheets, quizzes and videos with transcripts (https://www.thenational.academy/teachers) - designed and quality-assured by expert, subject specialist teachers in line with Oak's evidence-informed curriculum principles [7]. This content is openly licensed, on the Open Government Licence version 3.0 (OGL) which is compatible with Creative Commons by Attribution 4.0 (CC-BY) in line with UNESCO's Recommendation on OER [11].

This corpus of high-quality curricula content gives a valuable starting point for an AI-powered lesson planning tool that is free to use and accessible to UK teachers, as research has shown that providing

Second International Workshop on Generative AI for Learning Analytics, March 03, 2025, Dublin, Ireland
*Corresponding author.
hannah-beth.clark@thenational.academy (H. Clark)

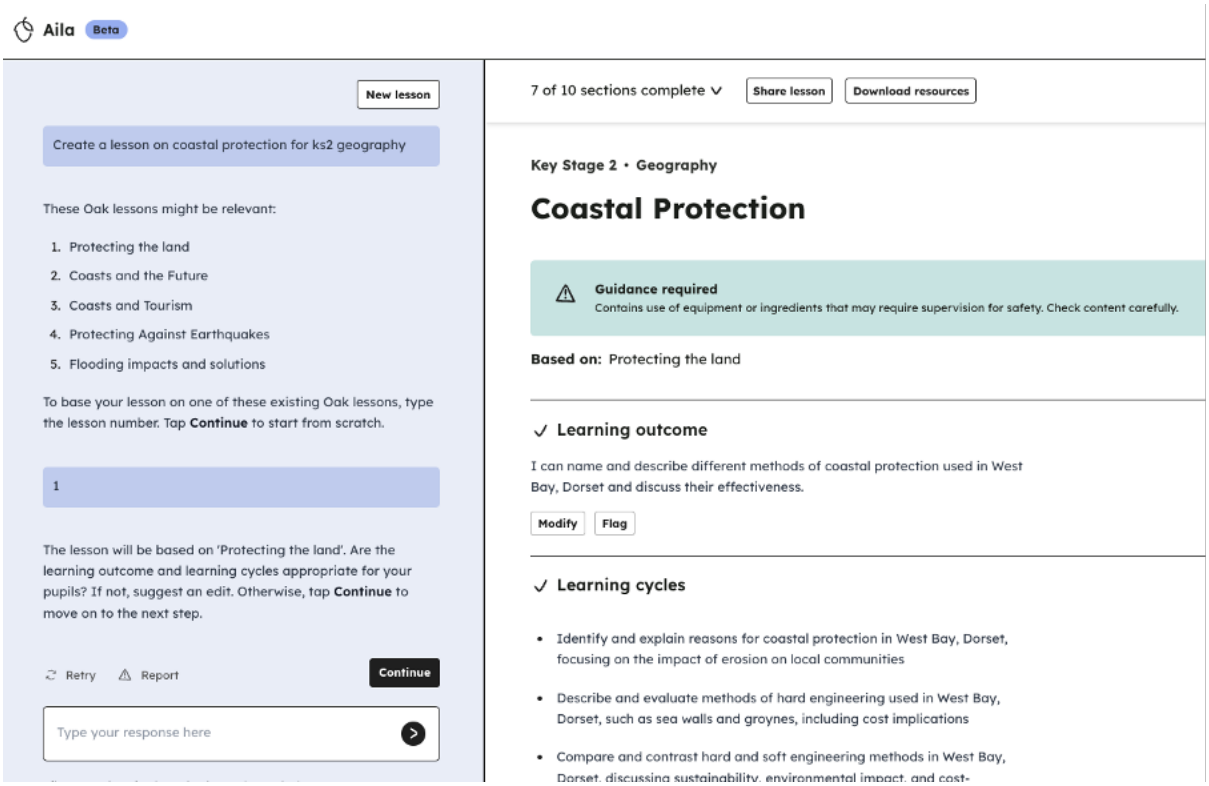

Figure 1: Lesson Planning Interface

generative AI models with a high-quality corpus in a retrieval database for use in RAG can improve accuracy from 67% to 92% [5]. In this paper, we describe our approach to designing Aila, our AI lesson assistant and the auto-evaluation agent built alongside to assess the accuracy, quality and safety of the lessons Aila produces. We also present empirical data from a case study to assess the effectiveness of this auto-evaluation agent.

2. System Design

Aila is designed to emulate the thought process of an experienced teacher as they plan a lesson. It is intentionally designed not to be a 'single-shot' tool that creates a lesson in one click, but instead supports teacher agency through enabling them to adapt and edit the lesson step-by-step to better suit their students (see Figure 1).

Our underlying content, alongside the codification of good practice in lesson design, has enabled us to use several techniques to raise the quality of Aila's outputs. These include retrieval augmented generation (RAG), to provide relevant context for the output [1] and more specifically content anchoring, to improve lesson quality by instructing the model to respond within the bounds of specified content (i.e. an existing Oak lesson) [6]; prompt engineering, to focus the response of the underlying Large Language Model (LLM) according to our codified definition of a high-quality lesson; and decision-making by the teacher at a granular level to act as the human in the loop [10, 12].

To enable us to understand the effectiveness of these techniques by evaluating Aila's outputs quickly and efficiently, we have built an auto-evaluation agent, using LLM as a Judge methodology [2], which is based on Oak's curriculum principles [7]. Each lesson is currently evaluated using a series of auto-evaluation prompts, assessing 24 quality and accuracy benchmarks, such as cultural bias, minimally different quiz answers or the progression of quiz difficulty (for the full list, see Appendix A). This has enabled us to evaluate the impact of the changes we make to improve Aila and compare the results, such as using different models as the underlying LLM, testing new versions of Aila before release, and

identifying particular areas for development, which is the focus of this paper.

3. Case Study

3.1. Task Description

Aila produces diverse educational resources, including lesson plans and classroom materials. We wanted to understand how closely aligned the auto-evaluation agent was with qualified teachers. To do this we first created a dataset of 2249 user-created Aila lessons, and 2736 lessons produced by Aila without user input or content anchoring (i.e. single shot), totalling 4985 lessons. The lessons were across all four key stages (i.e. for ages 5-16 years) and included maths, English, history, geography and science. The auto-evaluation model (gpt-4o-2024-08-06, temperature: 0.5) scored the lessons on 19 Likert criteria (using a 1-5 scale, see Figure 2) and 5 boolean criteria (true or false), each with their respective justifications.

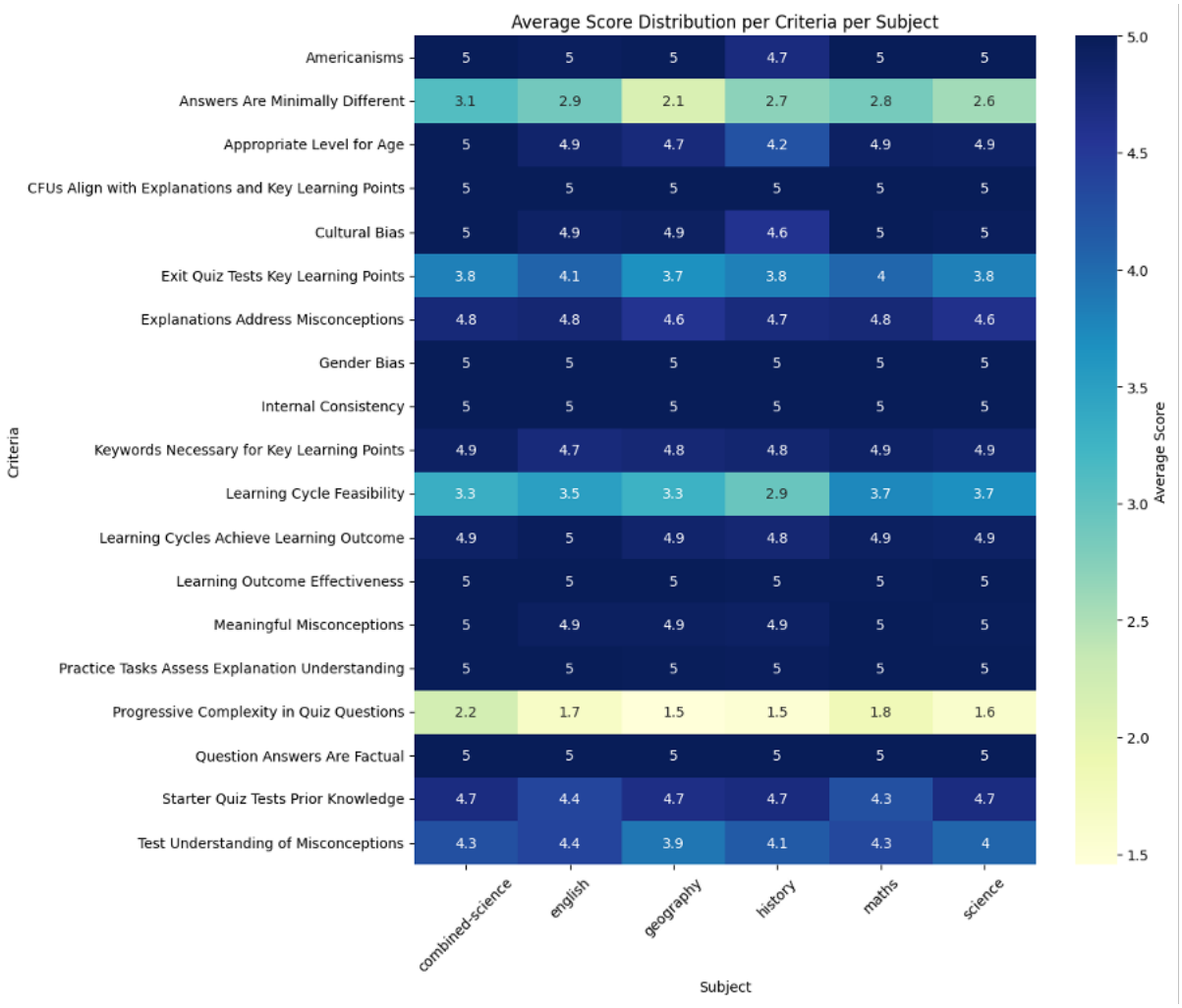

Figure 2: The Auto-Evaluation Tool Assesses Lessons Based on 19 Score-Based Criteria

3.2. Analysis

Our initial analysis focused on MCQs that teachers scored as 1, 3, and 5 to understand weak, average, and strong distractor quality, conducting a thematic analysis of the teachers' justifications for these scores. We limited our thematic analysis to these three categories to provide clear benchmarks for

quality assessment and to identify distinctive characteristics at each level of performance. We then identified exemplar MCQs to supplement the amended auto-evaluation prompts.

3.3. Results

3.3.1. What makes a generated distractor high or low-quality in relation to providing an appropriate level of challenge?

Appendix B summarises the key rating justification themes given by the human evaluators. The most common reason for distractors being low-quality was having the opposite sentiment to the correct answer (e.g. correct answer is a positive trait and the distractors are all negative traits). Other reasons included having a different grammatical structure to the correct answer, as well as the correct answer repeating words from the question, but the distractors not. For distractors to be high-quality they should fall into the same category as the correct answer, relate to a common theme, include common misconceptions and have a similar grammatical structure.

3.3.2. How well aligned were the auto-evaluation agent and the human evaluators?

Figure 3 highlights how the auto-evaluation agent was applying excessively strict criteria compared to the human evaluator, rating a large number of quiz questions as having low-quality distractors. It justified the low scores by claiming that the answer options were conceptually very different, thereby lacking the necessary challenge for the specified key stage. There was also an overemphasis on what was expected of students at the key stages, challenging deeper understanding.

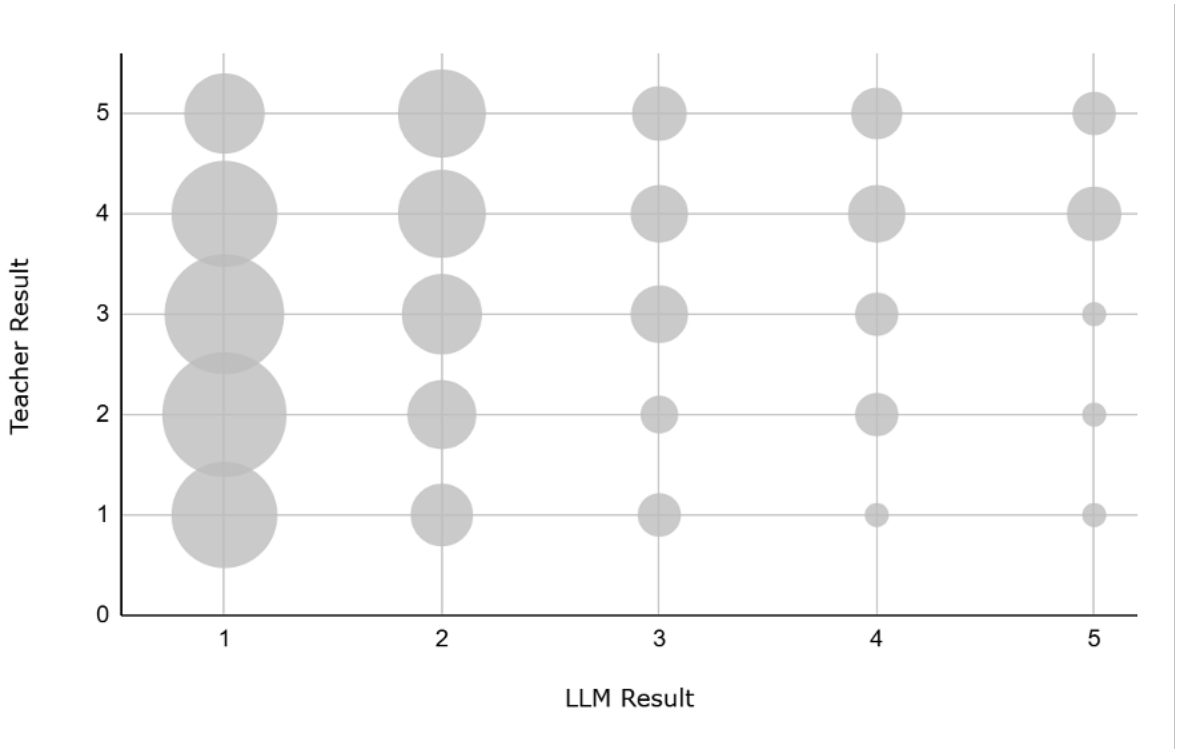

Figure 3: Paired scores of auto-evaluation and human evaluation, bubble size indicates the number of quiz questions.

We used the thematic analysis findings to update the prompt with additional guidance defining a high-quality distractor, and as a result, the auto-evaluation scores and human evaluation scores became more aligned (see Table 1). We calculated the Mean Squared Error (MSE) using the mean of the 10 scores given by the auto-evaluation per evaluation. The mean-based MSE decreased from 3.81 to 2.94 (p-value = 0.00679), which is statistically significant ($p < 0.05$). We also calculated several other evaluation

metrics, including the Quadratic Weighted Kappa (QWK), which showed an increase from 0.17 to 0.32, indicating a moderate to large and statistically significant improvement in agreement (see Appendix C).

Table 1
Count of differences between LLM and human scores before and after improving the prompt based on the thematic analysis findings.

LLM-Human score difference	Count (before)	Percentage (before)	Count (after)	Percentage (after)
0	61	19%	85	27%
1	97	31%	106	34%
2	80	26%	72	23%
3	58	19%	37	12%
4	17	5%	13	4%
Total lower scores by LLM	232	74%	194	62%
Total higher scores by LLM	20	6%	34	11%

4. Discussion

Through an illustrative case study, we have demonstrated the potential of using an auto-evaluation agent to drive improvement in the quality of AI-generated lessons and resources, as well as how the effectiveness of this agent can be improved by drawing on specific teaching expertise of human evaluators. Thematic analysis of rating justifications allowed us to codify what high and low-quality distractors looked like (with few-shot examples) and incorporate this information directly into the prompt, increasing the alignment with the human evaluators and driving improvements in the overall MCQ quality.

Incorporating the thematic analysis and corresponding representative examples for scores of 2 and 4 in future work could help reduce minor discrepancies by increasing granularity, especially in cases where scores are '1 away' from human evaluations. Absolute alignment is not necessarily the ultimate goal; the more important measure of success would be to see if the justifications the LLM gives alongside scores of 1, 3 and 5 are in line with the themes we found, providing consistent scoring according to these guidelines. Further thematic analysis would be required to establish this. Even after the changes, the LLM still scores lower than the human the majority of the time. This greater sensitivity is more beneficial than the alternative, as potential issues are more likely to be flagged and addressed.

There were also limitations to this work. We had a specific focus on answer differentiation and MCQs which could have implications for wider generalisability. Furthermore, due to time constraints, we weren't able to have multiple human evaluators for each question. Ideally, we would have an average human score per evaluation to deal with possible outliers. In future work, we could also consider weighting these responses according to the teacher's experience level, factoring in years of experience, teaching role and other metrics.

4.1. Recommendations

Aila has been designed specifically to support teachers in the UK with planning high-quality lessons and resources to reduce teacher workload and improve the quality of materials produced using AI. We hope by sharing what we have learned through this work it can also have an impact on other projects:

Having a base of high-quality OER has been integral to the quality of lessons produced by Aila. Our curriculum materials are aligned with the national curriculum for England, produced by expert teachers, available on an open government licence, and targeted at UK schools. For other organisations looking to develop tools within this space in other contexts, access to high-quality resources appropriate for their context will be imperative. We seek to enable this by making our OER resources available through a public API.

We had already done significant work codifying and exemplifying high-quality curriculum design. This provided invaluable input as the starting point for writing our prompt and, in turn, our evaluation tools. Deciding on your organisation's agreed-upon concept of "high-quality" is an important starting point before developing your tool, as this will be built into your prompt and evaluation work.

Using a cycle of comparative auto and human evaluations allowed us to iterate on the auto-evaluation prompt continuously and will ultimately also enable us to refine Aila's prompt. Once you have identified full lesson plans that achieve good scores aligned between evaluators through this iterative process these plans can subsequently be used to fine-tune generation models to output better-quality lesson plans [8].

4.2. Conclusion

We believe that auto-evaluation is a powerful tool for driving improvement in AI-produced content quickly and efficiently. We have focused specifically on a "quality" benchmark but we are also in the process of applying this approach to our "safety" benchmarks. The use of our auto-evaluation tool to evaluate different versions of Aila as we release them, comparisons of quality in how RAG is used, and the use of fine-tuning to develop the quality of our AI tools are further areas we plan to investigate. We also aim to use an improvement agent which will take feedback from our auto-evaluation agent to improve the quality of lesson content before it is displayed to users as well as suggest specific areas for users to check carefully or improve.

Declaration on Generative AI

Generative AI tools have not been used to support manuscript preparation.

References

[1] Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(16), 17754-17762.

[2] Chiang, C. H., & Lee, H. Y. (2023). A closer look into automatic evaluation using large language models. *arXiv preprint*, arXiv:2310.05657.

[3] Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence, 4*, 100118.

[4] D'Sa, J. L., & Wisbal-Dionaldo, M. L. (2017). Analysis of multiple choice questions: item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education, 9*(3).

[5] Government Social Research. (2024). Use Cases for Generative AI in Education: Building a proof of concept for Generative AI feedback and resource generation in education contexts [Technical report]. GOV.UK.

[6] Kommineni, V. K., König-Ries, B., & Samuek, S. (2024). From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. *arXiv preprint*, 2403.08345.

[7] McCrea, E. (2023). Our 6 principles guiding our approach to curriculum. Oak National Academy.

[8] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems, 35*, 27730-27744.

[9] Teacher Tapp. (2024). AI teachers, school exclusions and cutting workload. Teacher Tapp.

[10] Tsiakas, K., & Murray-Rust, D. (2022). Using human-in-the-loop and explainable AI to envisage new future work practices. *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 588-594.

[11] UNESCO. (2019). Recommendation on Open Educational Resources (OER) - Legal Affairs. UNESCO.

[12] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems, 135*, 364-381.

A. Full set of assessed quality and accuracy benchmarks

Table 2
Full set of assessed quality and accuracy benchmarks.

Prompt Criteria Title	Check Output Format	Relevant Lesson Plan Part	Criteria Group
Learning Cycle Feasibility	Likert	Key-Stage, Cycle-feedback, Cycle-practice, Cycle-explanations, Cycle-check	learning-cycles
Practice Tasks Assess Explanation Understanding	Likert	Cycle-practice, Cycle-explanations	learning-cycles
Keywords Necessary for Key Learning Points	Likert	Keywords, Key Learning Points	learning-outcomes
CFUs Align with Explanations and Key Learning Points	Likert	All cycles	learning-outcomes
Learning Cycles Achieve Learning Outcome'	Likert	Learning Outcome, Learning Cycle	learning-outcomes
Learning Outcome Effectiveness	Likert	Learning Outcome	learning-outcomes
Explanations Address Misconceptions	Likert	Cycle-explanations, Misconceptions	misconceptions
Meaningful Misconceptions	Likert	Misconceptions, Topic	misconceptions
Test Understanding of Misconceptions	Likert	Exit Quiz, Cycle-check, Misconceptions	misconceptions
Question Answers Are Factual	Likert	Whole Lesson	lesson-quality
Internal Consistency	Likert	Whole Lesson	lesson-quality
Appropriate Level for Age	Likert	Whole Lesson, Key-Stage	bias
Americanisms	Likert	Whole Lesson	bias
Cultural Bias	Likert	Whole Lesson	bias
Gender Bias	Likert	Whole Lesson	bias
Exit Quiz Tests Key Learning Points	Likert	Exit Quiz, Key Learning Points	quizzes
Starter Quiz Tests Prior Knowledge	Likert	Starter Quiz, Prior Knowledge	quizzes
Answers Are Minimally Different	Likert	Starter Quiz, Exit Quiz	quizzes
Progressive Complexity in quiz Questions	Likert	Exit Quiz	quizzes
Learning Cycles Increase in Challenge	Boolean	Learning Cycles	learning-cycles
No Negative Phrasing in Quiz Questions	Boolean	Starter Quiz, Exit Quiz	quizzes
Repeated Questions in Quizzes	Boolean	Exit Quiz, Starter Quiz	quizzes
Starter Quiz does not Rest Lesson Content	Boolean	Starter Quiz, Learning Cycles, Learning Outcome, Key Learning Points, Prior Knowledge	quizzes
Exit Quiz Contains Vocabulary Question	Boolean	Exit Quiz, Keywords	quizzes

B. Summary of thematic analysis

Table 3
Thematic Analysis of Quiz Questions

Mean human score	Theme	Frequency	Example
1.5	Distractors have an opposite sentiment to the right answer or the question	27	Which of these is a positive impact of TNCs in the food industry? **Creating jobs in developing countries**, Making local foods more expensive., Eliminating smaller companies., Reducing dietary variety
1.9	The correct answer is structurally different	21	What is the Atacama Desert known for? **Being one of the driest places on Earth**, Its large rainforest, Its snowy mountains, Its tropical beaches
1.7	The right answer repeats words from the question	11	What is required to simplify an algebraic fraction by factorisation? **Factorise the quadratic expressions in the numerator and denominator**. Multiply the numerator and denominator by a common factor., Add the expressions in the numerator and denominator., Subtract the denominator from the numerator.
3.0	One of the distractors is semantically different	21	Which period did William Wordsworth belong to? **Romantic** Victorian, Elizabethan, Modernist
3.0	No obvious mistakes in the quiz but it lacks sufficient challenge	17	What does TNC stand for? **Transnational Corporation**, Total National Company, Trade Negotiation Committee, Territorial Network Corporation
2.6	Distractors do not address typical misconceptions	10	If the probability of an event is 0.5, what is the probability of the opposite event? **0.5**, 1, 0, It cannot be determined
3.0	Two options are semantically very different to the other two options	6	What is an astronaut? **A scientist trained to go into space**, A space tour guide, A pilot who flies airplanes, A doctor who treats illnesses
3.0	Answer options fall into different categories	4	Which term describes medieval stories of knights and romance? **Chivalry**, Allegory, Satire, Fable
5.0	All quiz answers fall into the same category	14	Which material is not commonly used to make shell structures? **Wood**, Cardboard, Plastic, Metal
5.0	All answers relate to a common theme	10	Which term describes the complete outer edge of a circle? **circumference**, radius, diameter, segment
5.0	Distractors include common misconceptions	8	What was the primary crop grown on plantations in early Virginia? **Tobacco**, Cotton, Sugar, Rice
5.0	All answers are structurally similar	4	How does Wordsworth portray nature in the poem? **As beautiful and restorative**, As mundane and uninteresting, As artificial and man-made, As oppressive and confining

C. Evaluation metrics

Table 4
Results before the thematic analysis

Score	Precision	Recall	F1	Accuracy	QWK
1	0.19	0.78	0.30	0.19	0.17
2	0.14	0.16	0.15	–	–
3	0.30	0.08	0.13	–	–
4	0.43	0.08	0.13	–	–
5	0.29	0.04	0.07	–	–

Table 5
Results after the thematic analysis

Score	Precision	Recall	F1	Accuracy	QWK
1	0.22	0.78	0.35	0.27	0.32
2	0.23	0.18	0.20	–	–
3	0.24	0.11	0.15	–	–
4	0.42	0.27	0.33	–	–
5	0.50	0.18	0.26	–	–