Assessing Educational Quality: Comparative Analysis of Crowdsourced, Expert, and AI-Driven Rubric Applications

Steven Moore, Norman Bier, John Stamper
Carnegie Mellon University
StevenJamesMoore@gmail.com, nbier@cmu.edu, jstamper@cmu.edu

Abstract

Exposing students to low-quality assessments such as multiple-choice questions (MCQs) and short answer questions (SAQs) is detrimental to their learning, making it essential to accurately evaluate these assessments. Existing evaluation methods are often challenging to scale and fail to consider their pedagogical value within course materials. Online crowds offer a scalable and cost-effective source of intelligence, but often lack necessary domain expertise. Advancements in Large Language Models (LLMs) offer automation and scalability, but may also lack precise domain knowledge. To explore these trade-offs, we compare the effectiveness and reliability of crowdsourced and LLM-based methods for assessing the quality of 30 MCQs and SAQs across six educational domains using two standardized evaluation rubrics. We analyzed the performance of 84 crowdworkers from Amazon's Mechanical Turk and Prolific, comparing their quality evaluations to those made by the three LLMs: GPT-4, Gemini 1.5 Pro, and Claude 3 Opus. We found that crowdworkers on Prolific consistently delivered the highest-quality assessments, and GPT-4 emerged as the most effective LLM for this task. Our study reveals that while traditional crowdsourced methods often yield more accurate assessments, LLMs can match this accuracy in specific evaluative criteria. These results provide evidence for a hybrid approach to educational content evaluation, integrating the scalability of AI with the nuanced judgment of humans. We offer feasibility considerations in using AI to supplement human judgment in educational assessment.

Introduction

Multiple-choice questions (MCQs) and short answer questions (SAQs) are widely utilized in educational assessments due to their versatility across various learning environments (Butler 2018; Lu et al. 2021). Despite their popularity, creating high-quality and reliable educational assessments is challenging, often requiring significant time and domain-specific expertise (Cochran et al. 2022; Haladyna, Downing, and Rodriguez 2002). Existing tools and methods for crafting and evaluating these questions are not without their issues, capable of producing questions with inherent flaws that are potentially detrimental to their pedagogical value. These flaws can persist in widely used question datasets and across online courses, hindering the student learning process (Costello, J. Holland, and Kirwan 2018b; Rush, Rankin, and White 2016). The gold standard for identifying and correcting these issues traditionally involves expert human judgment (Kamalloo et al. 2023). Automated evaluation methods, although less subjective, typically depend on extensive student performance data or focus on superficial metrics like readability, which do not fully capture the educational effectiveness of the questions or correlate with human judgment (Azevedo, Oliveira, and Beites 2019).

Despite the recognized need for human expertise in evaluating the quality of educational assessments, relying solely on such input limits the scalability and efficiency of the process. Crowdsourcing and learnersourcing offer potential solutions by leveraging collective human intelligence on a larger scale, though these methods often involve participants with less expertise (Singh, Brooks, and Doroudi 2022; Sun et al. 2022). Moreover, recent developments in Large Language Models (LLMs) suggest that AI could mimic human-like judgment for certain educational tasks, offering a scalable approach for assessing question quality (Addlesee et al. 2023; Markel et al. 2023).

In response to these challenges, this study compares the effectiveness of multiple crowdsourcing strategies with LLM-based methods for evaluating the quality of 30 MCQs and SAQs across six domains. These evaluations employ two standardized and validated rubrics, examining the assessments' pedagogical validity. We conducted two distinct crowdsourcing tasks, one for MCQs and another for SAQs, to see how well novice contributors could apply these rubrics. Concurrently, we utilized three state-of-the-art LLMs to automate the same evaluation process. By analyzing the wisdom of the crowds (Kremer, Mansour, and Perry 2014) this research assesses how closely the majority responses from crowdsourced evaluations align with those generated by LLMs and verified by subject matter experts. This study investigates two primary research questions: How do the effectiveness and accuracy of rubric applications by crowdworkers, experts, and AI models compare when assessing educational content (**RQ1**)? How consistent and reliable are quality assessments of MCQs and SAQs within crowdsourced and LLM methods (**RQ2**)?

Through the investigation of these research questions, this work makes the following contributions: First, it demonstrates the comparative effectiveness and accuracy of rubric applications by crowdsourced workers, experts, and LLMs in evaluating the quality of educational assessments. Second, it provides a detailed analysis of the consistency and reliability of quality evaluations for MCQs and SAQs, highlighting critical trade-offs. Finally, it provides insights into the integration of LLMs in the educational quality evaluation process, proposing a potential hybrid approach that leverages both human expertise and AI to enhance the quality and reliability of educational assessments.

Related Work

Educational Rubrics

Human evaluation is traditionally considered the gold standard for accurately assessing the quality of educational content (Amidei, Piwek, and Willis 2018a; Mulla and Gharpure 2023). Despite being the benchmark metric, human evaluation often relies on subjective metrics such as difficulty or acceptability, which are based largely on the evaluator's interpretation (Bates et al. 2014; van der Lee et al. 2021). An objective alternative, the Item-Writing Flaws (IWF) rubric, has been effectively validated and used for MCQ quality evaluation (Tarrant et al. 2006). This rubric includes 19 criteria designed to evaluate MCQs in any academic subject, focusing on pedagogical aspects beyond mere readability and superficial features. The effectiveness of it has been demonstrated across various disciplines, ranging from STEM to the humanities, proving its utility in evaluating high-quality educational MCQs (Breakall, Randles, and Tasker 2019; Pate and Caldwell 2014; Rush et al. 2016; Tarrant and Ware 2008). Similarly, for SAQs, a 9-item rubric has been validated and employed across prior research (Horbach et al. 2020; Moore et al. 2022; Steuer et al. 2021). This SAQ rubric assesses both linguistic and pedagogical qualities, with certain criteria requiring evaluators to have domain-specific knowledge. In this study, we use the 19-criteria IWF rubric and the 9-criteria SAQ rubric to assess question quality, providing comprehensive standardization.

Crowdsourcing in Education

Crowdsourcing has increasingly been utilized in educational settings to enhance the generation of questions, provide feedback, and improve question content (Abdi, Khosravi, and Sadiq 2020; Khosravi, Kitto, and Williams 2019). Educational crowdsourcing tasks frequently demand domain knowledge, which crowdworkers may not inherently possess (Kobren et al. 2014). To address this issue, previous research has provided crowdworkers with clearly defined rubrics to guide their work (Doroudi, Kamar, and Brunskill 2019; Labutov and Studer 2017). The use of such rubrics has proven effective in enhancing the quality of responses, reducing the time crowdworkers spend on tasks, and mitigating challenges associated with a lack of domain expertise (Yuan et al. 2016). For example, one study provided crowdworkers, who lacked background knowledge in teaching and writing, with a rubric for assessing student writing artifacts (Ahn et al. 2021). The study found that the crowdsourced evaluations closely aligned with expert assessments, demonstrating substantial agreement. In this study, rubrics are employed to guide crowdworkers to assess the quality of educational questions effectively.

Automating Educational Quality Evaluation

Automated question quality evaluation in education frequently employs metrics such as BLEU, METEOR, and ROUGE (Mulla and Gharpure 2023). These metrics measure the similarity of responses to a predefined gold standard, but they do not account for the educational value or effectiveness in assessing student knowledge (Moon et al. 2022). Moreover, previous research has shown that these metrics often do not align well with human evaluations, highlighting a gap in their ability to measure quality accurately (Gao et al. 2024; van der Lee et al. 2021). In response, recent studies have explored the use of LLMs to assess the quality of educational content more effectively due to their access to domain-specific information. By integrating rubrics to guide the LLMs, researchers have achieved promising results, with evaluations that are comparable to those performed by human experts for specific rubric criteria for the task of essay scoring (Stahl et al. 2024; Yavuz, Çelik, and Yavaş Çelik 2024). This approach not only enhances the accuracy of automated assessments, but also ensures that the evaluations consider the pedagogical significance of the questions.

Methodology

To explore the effectiveness and trade-offs between crowdsourced and programmatic LLM-based methods in assessing the quality of educational questions, we conducted a comparative study across two types of questions, multiple-choice questions (MCQs) and short answer questions (SAQs), spanning six subject areas. Our study comprised two experiments: the first evaluated the quality of MCQs using the 19-criteria IWF rubric, applied by various crowdworkers and multiple LLM-based programmatic methods; the second experiment involved a similar evaluation of SAQs using a 9-item rubric. In total, 30 questions were evaluated, 15 MCQs and 15 SAQs, from distinct domains within mathematics, science, and the humanities. All the questions were purely text-based, with no accompanying images or formulas. The 15 MCQs used in this research were sourced from a previous study, where the IWF rubric had already been applied (Costello, J. Holland, and Kirwan 2018a). These MCQs were extracted from introductory online courses in Philosophy, Statistics, and Chemistry.

The five SAQs related to Chemistry were obtained from a separate study involving an online introductory Chemistry course (Moore et al. 2022). We selected five SAQs each from online Calculus and Team Collaboration courses at a university on the U.S. East Coast. The Team Collaboration course covers communication, teamwork, and conflict management. These questions were selected by two domain experts, who identified potential flaws within them. For all 15 of the SAQs, the experts applied the 9-item SAQ rubric to evaluate these questions. To assess the consistency of their evaluations, we calculated the inter-rater reliability using Cohen's Kappa (McHugh 2012). The overall Cohen's Kappa score was 0.79, indicating substantial agreement between the raters across the entire rubric. Further details about these questions are available in Table 1.

Domain	Type	Number	Number of Flaws
Philosophy	MCQ	5	10
Statistics	MCQ	5	11
Chemistry	MCQ	5	10
Team Collaboration	SAQ	5	12
Calculus	SAQ	5	6
Chemistry	SAQ	5	11

Table 1: Information about the 30 questions.

Item-Writing Flaws Rubric

To evaluate the MCQs, we engaged both crowdworkers and LLMs to apply the IWF rubric. The IWF rubric encompasses 19 criteria specifically designed to assess the quality of educational MCQs. This version of the rubric has been extensively used and validated in previous research, particularly within STEM fields (Breakall et al. 2019; Rush et al. 2016). The criteria cover various aspects of the questions, including the question text, answer choices, and the correct option, ensuring a comprehensive evaluation of each component. While expertise is not required to apply this rubric, certain criteria, such as identifying implausible distractors or logical cues, may benefit from domain knowledge as well as an understanding of assessment creation. A detailed list of the 19 IWFs and their definitions is provided in Table 2.

Item-Writing Flaw	Definitions
Absolute Terms	Use of definitive words like "always" or "never" that can make a statement true or false.
All of the Above	Inclusion of an option that suggests selecting all previous options.
Ambiguous Information	Unclear or vague content that can lead to multiple interpretations.
Convergence Cues	Clues within the question or options that guide test-takers to the correct answer.
Logical Cues	Answer choices that can be deduced logically rather than through knowledge of the subject.
Complex or K-type	Use of complex formats like multiple true-false questions within a single item.
Fill in the Blank	Questions requiring a missing word or phrase, which can be too open-ended.
Grammatical Cues	Grammatical inconsistencies between the stem and the correct answer can act as a hint.
Gratuitous Information	Unnecessary details that do not contribute to the question, potentially distracting the test-taker.
Implausible Distractors	Option choices that are obviously incorrect, making the question too easy.
Longest Correct	The correct answer is noticeably longer than the distractors.
Lost Sequence	Options that are not presented in a logical or sequential order, causing confusion.
More than One Correct	Multiple correct answers when only one is expected, causing ambiguity.
Negative Wording	Use of negative phrases like "Which of the following is NOT..." that can confuse test-takers.
None of the Above	Including an option that invalidates all other choices, which can be misleading.
True or False	Avoid simplistic questions using true or false, as they reduce the depth of assessment.
Unfocused Stem	The question stem is not clear or concise, leading to confusion about what is being asked.
Vague Terms	Use of unclear or imprecise terms that can be interpreted in multiple ways.
Word Repeats	Repetition of words or phrases in the stem and the correct answer, providing unintended hints.

Table 2: Definitions for each of the 19 criteria used to identify common item writing flaws in educational MCQs.

Short Answer Question Rubric

To evaluate the SAQs, we employed both crowdworkers and LLMs to apply a 9-item rubric. The rubric was from two previous studies that used a version of it for evaluating

STEM questions (Horbach et al. 2020; Steuer et al. 2021). We adjusted the rubric by combining elements from both studies to minimize the inclusion of overly subjective criteria. However, unlike the more objective IWF rubric, this SAQ rubric still contains criteria that can be interpreted differently. The final version of the SAQ rubric used in this study is detailed in Table 3. It lists the criteria labels along with corresponding yes-or-no questions that assess whether each criterion is met or violated. It is important to note that the answer to the SAQ or any other associated metadata is not required for applying this rubric's criteria during the evaluation process. Like the IWF rubric, applying this rubric may be easier for evaluators with domain knowledge, particularly for criteria that specify [*specific domain*].

Criteria	Evaluation Questions
Understandable	If you were a student in a [specific domain] course, could you clearly understand this question without additional explanations?
Domain Related	Is the question related to [specific domain]?
Grammatical	Is the question grammatically correct and free of language errors?
Focus	Is the question specific and focused on a single concept or topic?
Conciseness	Is the question concise and free of unnecessary information?
Fairness	Is the question culturally neutral and free from any biases that might disadvantage any group of students?
Cognitive Level	Does the question require students to apply higher-order thinking skills rather than simply recalling facts?
Central	Is being able to answer the question important for understanding the topics covered by a course in [specific domain]?
Would You Use It	If you were a teacher working with content related to this question in your course, would you include this question in the course?

Table 3: Questions used for evaluating the quality of SAQs based on the 9-item rubric.

Participants

We recruited participants using two popular crowdsourcing platforms, Amazon's Mechanical Turk (MTurk) and Prolific (Douglas, Ewell, and Brauer 2023). On each platform, participants received identical instructions for the task, which involved completing a survey. In this survey, they evaluated five questions at a time, applying the appropriate rubric based on the type of question being assessed.

To evaluate the 15 MCQs using the 19-criteria IWF rubric, we adapted it into yes-or-no questions for crowdworkers to assess whether each MCQ violated specific criteria. Participants were presented with each MCQ, which included the question text and four answer choices. They were informed that the first choice, option A, was the correct answer. This process is conducted on five distinct MCQs drawn from introductory courses in Philosophy, Statistics, or Chemistry. The crowdworkers evaluate each of the 19 criteria for one MCQ before moving on to the next, completing evaluations for a total of five MCQs.

Similar to the IWF task for MCQs, the evaluation process for SAQs involved applying the 9-criteria SAQ rubric to each question, structured as a series of nine yes-or-no questions. Crowdworkers assessed each SAQ individually, completing evaluations for five SAQs sequentially. Each set of SAQs covered one of the three domains used in this experiment: Team Collaboration, Calculus, or Chemistry.

We recruited two distinct groups of participants: novice crowdworkers from MTurk and those with some domain expertise from Prolific. All participants were over 18 years old, self-reported as expert or native English speakers, and were compensated at a rate of at least $18 per hour for their time. The tasks were designed to be efficient: the average completion time for the MCQ task was 14 minutes and 36 seconds, while the SAQ task took an average of 9 minutes and 2 seconds. Upon completing the task, participants were asked to self-report their understanding using a five-point Likert scale and to provide any written feedback. All participants from both platforms reported scores of 4 or 5, indicating a high level of full understanding of the task.

Amazon's Mechanical Turk We utilized MTurk to recruit 11 unique crowdworkers for each of the six question evaluation tasks, totaling 66 participants. To ensure high-quality contributions without severely limiting our participant pool, we established qualifications requiring that each crowdworker had an overall approval rate greater than 95% before they could participate in our study. These participants were considered novice, as none reported having professional experience in the domains of the questions or in education more broadly.

Prolific We recruited 18 unique crowdworkers from Prolific, assigning three crowdworkers to each of the six tasks. Each participant possessed at least a bachelor's degree in the domain relevant to their assigned questions. For example, the crowdworkers evaluating the five Calculus SAQs held degrees in mathematics. This group was considered more advanced and knowledgeable than those from MTurk, owing to their specialized educational backgrounds. Previous research has shown that Prolific generally attracts a higher-skilled audience capable of delivering superior results (Douglas et al. 2023). Due to these factors and associated cost considerations, we decided to limit the number of Prolific crowdworkers to three per task, with each group evaluating a set of five questions from one of the six domains.

Application of Large Language Models

We employed three LLMs, GPT-4, Gemini 1.5 Pro (Gemini), and Claude 3 Opus (Claude), to programmatically apply the two rubrics to our question set (Kevian et al. 2024). These three models were chosen for their strong performance benchmarks, widespread popularity at the time, and easy API access. For the automated application of the IWF rubric, we utilized an established automated method that applies various NLP techniques tailored to each of the 19 criteria (Moore et al. 2023b). This method has been previously applied in several studies involving MCQs in domains such as Biology and Algebra (Arif, Asthana, and Collins-Thompson 2024; Hwang et al. 2023; Moore et al. 2024). While this automated method incorporates the use of an LLM, for our current study, we varied which LLM was employed in each evaluation cycle to assess their relative effectiveness.

For the SAQs, our approach mirrored that of previous studies which have successfully used LLMs to apply rubric criteria to educational content (Jury et al. 2024; Yavuz et al. 2024). Adopting the LLM prompting strategy of having it assume the role of an expert, we assigned the LLMs the role of an experienced instructor tasked with evaluating the quality of educational content (Liu et al. 2024). Given that the 9-item SAQ rubric consists of yes-or-no questions corresponding to each criterion, we used these questions as prompts for the LLMs, inputting both the rubric question and the text of the SAQ for evaluation.

The total cost and time required to evaluate all 15 MCQs and 15 SAQs using these methods are detailed in Table 4. We used a single iteration of LLM prompting for each set of questions, without running multiple iterations or combining outputs, to align with methods used in previous research.

Method	Type	Cost (USD)	Time (seconds)
GPT-4	MCQ	0.21	28
GPT-4	SAQ	0.72	77
Claude 3 Opus	MCQ	0.13	504
Claude 3 Opus	SAQ	0.24	1557
Gemini 1.5 Pro	MCQ	0.04	33
Gemini 1.5 Pro	SAQ	0.12	79

Table 4: The cost and time of GPT-4, Gemini 1.5 Pro, and Claude 3 Opus in applying the IWF and SAQ rubrics.

Data Analysis

For each evaluation task, we assessed the crowdworkers' ability to effectively apply the specified criterion from each rubric using a consensus-based approach. Specifically, we adopted the majority response as the representative outcome for each criterion. For example, in the Calculus MCQ task, if six out of eleven crowdworkers indicated that a question violated the first criterion of the IWF rubric, this majority view was taken as the crowd's consensus. This method, often referred to as the “wisdom of the crowd", is a widely used technique for aggregating responses from crowdsourcing platforms (Kremer et al. 2014).

For our comparison of accuracy between the crowdsourced and LLM methods, we referred to the human evaluations within our dataset. We addressed this multi-label classification challenge using the Exact Match ratio, which requires correct identification of all labels for a question to be considered a match, and the Hamming Loss, which calculates the average proportion of incorrect labels, providing detailed insights into our classification's holistic accuracy (Pi et al. 2020). Performance was further assessed using the Micro F1 score for each criterion, which combines precision (the accuracy of positive predictions) and recall (the completeness of positive predictions) to deliver a measure of each method's effectiveness in accurately classifying each specific criterion of the rubrics (Zhang, Wang, and Zhao 2015). A high Micro F1 score indicates both high precision and high recall, indicating effective identification of criteria with minimal false positives or negatives.

Additionally, we evaluated the Macro F1 score, which averages the F1 scores computed for each criterion independently, showing how uniformly the method performs across diverse categories without being influenced by the frequency of each criterion (Zhang et al. 2015). Finally, we utilized the Jaccard Index as another metric (Fletcher and Islam 2018). This index measures the intersection over the union of the predicted and actual labels at an aggregate level, offering a direct indicator of the overlap between the two sets. This metric is valuable for assessing the overall effectiveness of the classification in scenarios where accurate positive identifications are essential output integrity.

Results

RQ1: Crowdsourcing Outperformed the LLMs

MCQ Quality The 19 IWF criteria were applied to all 15 MCQs, resulting in a total of 285 classifications. The evaluation metrics for our five assessment methods, calculated by comparing them to the ground truth MCQ labels previously provided by two experts, are presented in Table 5.

For raw accuracy, MTurk shows the highest exact match ratio at 33%, indicating it has the highest proportion of correct predictions, exactly matching expert labels for five of the fifteen questions. It also has the lowest Hamming Loss at 7%, indicating a small amount of misclassification regarding the flaws. Due to having the most precise prediction with fewest incorrect labels, it is the best performing method for the IWFs. However, Prolific excels in the other three metrics, demonstrating the best balance of precision and recall across all 19 IWF criteria.

In comparison, the three automated methods perform worse than the crowdsourcing methods of MTurk and Prolific. Among the automated methods, GPT-4 performs the best, achieving moderate performance across all evaluation metrics. The poorest performer overall is Gemini, which has the lowest scores across all five evaluation metrics.

Method	Exact Match	Hamming Loss	Micro F1	Macro F1	Jaccard Index
MTurk	**0.333**	**0.070**	0.643	0.385	0.474
Prolific	0.200	0.081	**0.667**	**0.535**	**0.500**
GPT-4	0.200	0.091	0.567	0.370	0.395
Gemini	0.067	0.140	0.444	0.316	0.286
Claude	0.133	0.105	0.500	0.319	0.333

Table 5: Performance of 5 methods at applying the 19-criteria IWF rubric to 15 educational MCQs.

SAQ Quality The 9-item rubric was applied to all 15 SAQs, resulting in a total of 135 classifications. The evaluation metrics for our five assessment methods, calculated by comparing them to the ground truth SAQ labels provided by two experts, are presented in Table 6.

Method	Exact Match	Hamming Loss	Micro F1	Macro F1	Jaccard Index
MTurk	0.200	0.193	0.886	0.875	0.795
Prolific	0.200	**0.163**	**0.897**	0.879	**0.814**
GPT-4	**0.267**	0.185	0.892	**0.882**	0.805
Gemini	0.133	0.193	0.876	0.864	0.779
Claude	0.133	0.244	0.841	0.827	0.725

Table 6: Performance of 5 methods on applying the 9-criteria rubric for evaluating educational SAQs.

While GPT-4 achieves the highest exact match ratio, successfully classifying the most SAQs accurately across all 9 criteria, it is not the best overall performer. Prolific stands out by achieving high evaluation metrics, particularly in Micro F1 and Jaccard Index, which indicate a strong balance of precision and recall. Additionally, Prolific has the lowest Hamming Loss of all methods, indicating it is the most accurate in labeling. Similar to its performance in the MCQ evaluation, MTurk also performs well, achieving metrics only slightly lower than Prolific.

The remaining two automated methods, Gemini and Claude, performed poorly by comparison. Claude achieved the worst results across all evaluation metrics. This shows that while some automated methods can be effective, there is significant variation in their performance. Overall, Prolific emerges as the most reliable method for evaluating the quality of educational SAQs, combining high precision and recall with the lowest rate of labeling errors.

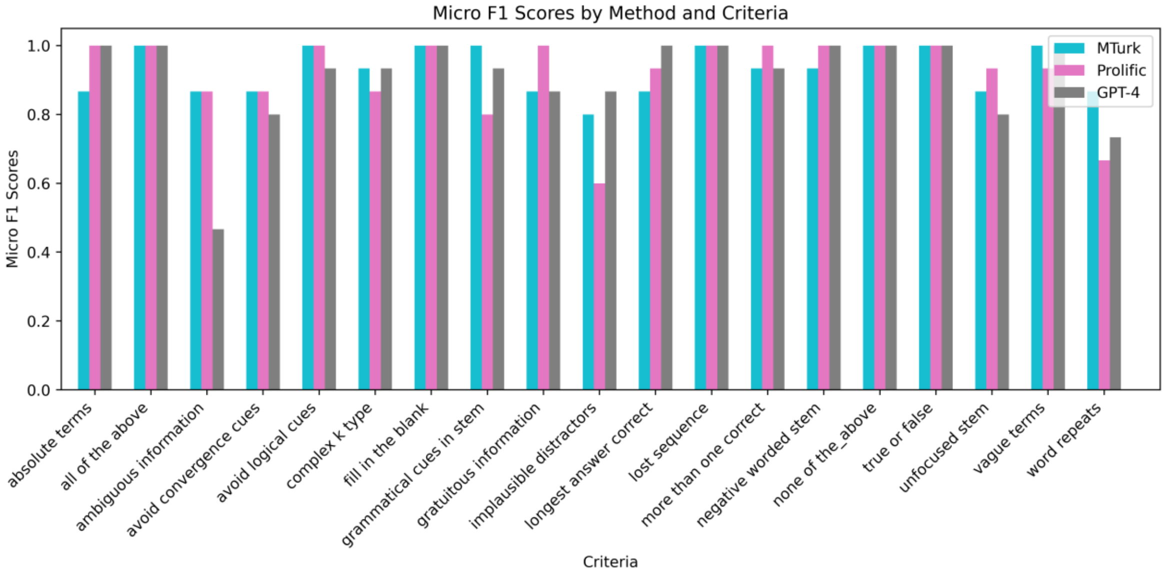

Figure 1: Comparison of Micro F1 scores across 19 IWF criteria for MTurk, Prolific, and GPT-4, illustrating the performance of the three top-performing methods in evaluating MCQs.

RQ2: Evaluating Method Trade-Offs

While much of the evaluation metrics focused on exact match and Hamming Loss, these do not provide a holistic picture. Exact match is strict and can be skewed by a few incorrect predictions, while Hamming Loss offers only a broad error rate. To provide a more comprehensive evaluation, we use the Micro F1 score, which considers both precision and recall, enabling a more accurate and realistic assessment of each method's effectiveness (Zhang et al. 2015).

IWF Performance We present a comparison of Micro F1 scores for the top three performing methods, MTurk, Prolific, and GPT-4, across all 19 criteria from the IWF rubric, as shown in Figure 1.

Less subjective criteria, which are simple enough to be addressed through programmatic string matching, such as *absolute terms*, *all of the above*, and *fill in the blank*, perform highly across all three methods. In contrast, more subjective measures that might be influenced by domain knowledge or instructional design preferences posed a challenge for the three methods. For instance, *ambiguous information* is one of the lower-scoring criteria, especially for GPT-4, indicating a difficulty in handling ambiguity in text. Similarly, *implausible distractors* present a challenge for all three methods, although GPT-4 performs the best in this area despite it requiring domain knowledge. Separating the evaluation by criteria further demonstrates that Prolific consistently achieves the highest performance. Notably, there are multiple criteria where all three methods perform at the highest level, achieving a Micro F1 score of 1.

SAQ Rubric Performance We present a comparison of Micro F1 scores for the top three performing methods, MTurk, Prolific, and GPT-4, across all 9 criteria from the SAQ rubric, as shown in Figure 2.

The most subjective criterion, *would you use it*, posed a challenge, as even the experts who reviewed these questions to create our ground truth struggled with this criterion, as it is purely subjective and influenced by many factors. Less subjective criteria, such as the *conciseness* of the text, also had poor performance across all methods.

The *understandable* criteria achieved high performance despite its potential subjectivity and influence from domain knowledge. Compared to the two crowdsourcing methods, GPT-4 achieved superior performance in *cognitive level* and *grammatical* criteria. Machines are typically good at these tasks, as cognitive level can be partly determined by verb usage (Assaly and Smadi 2015), and grammatical correctness has been a significant focus of LLMs and NLP work (Yavuz et al. 2024). Even though Prolific generally achieved the highest or tied for the highest performance on each criterion, GPT-4 performed quite close to it.

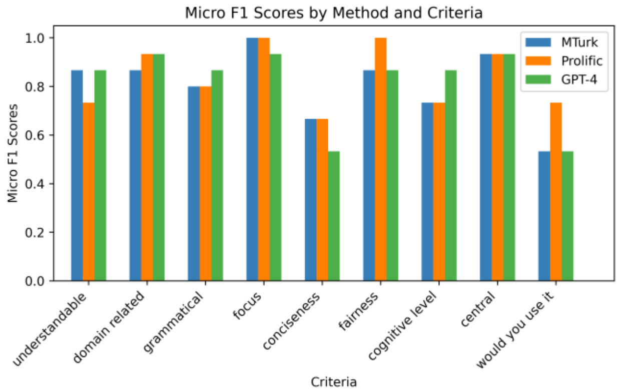

Figure 2: Comparison of Micro F1 scores for evaluating SAQs across 9 criteria by the three methods.

Discussion

In this study, we evaluated 30 questions, five from each of six distinct domains, using two crowdsourcing platforms and three state-of-the-art LLMs to apply two different rubrics. The results indicate that while the human-involved crowdsourcing methods generally outperformed the automated approaches, the LLMs performed comparably well on many criteria and even exceeded human performance on some. Across both types of questions at least one method, automated or crowdsourced, achieved perfect or near-perfect classification for a given criteria in alignment with human expert labels. These findings support the potential for a hybrid approach where human expertise is utilized primarily for the most challenging criteria, while AI handles the more straightforward tasks.

Method Evaluation We observed that the crowdsourcing methods outperformed the programmatic methods across the two tasks. Specifically, MTurk was the top performer for IWFs and Prolific excelled in the SAQ rubric evaluation. MTurk's workers typically possess less domain knowledge, a factor we controlled by selecting participants with relevant expertise on Prolific (Douglas et al. 2023; Moore et al. 2023a). The detailed and less subjective nature of the IWF rubric likely aided MTurk workers by providing sufficient guidance, despite their varied knowledge levels. For the SAQ task, Prolific's superior performance is attributable to our targeted recruitment of individuals with relevant academic qualifications. This was crucial since the SAQ task's criteria are inherently more subjective and knowledge-intensive (Steuer et al. 2021).

Regarding automated methods under the same research question, they generally underperformed when compared to the crowdsourced approaches. Notably, the latest iterations of Gemini 1.5 Pro and Claude 3 Opus were less effective than GPT-4, while also taking longer to complete (see Table

4). However, across all evaluation metrics, crowdsourced methods consistently outperformed automated ones. Despite the challenges of these 19-item and 9-item multi-label classification tasks, where achieving an exact match required correct labeling of each item, all methods managed to maintain a low hamming loss rate, demonstrating a base level of competency in handling these complex tasks.

Performance Variability In applying the IWF rubric, both crowdsourcing methods either matched or outperformed the automated GPT-4 process, with the notable exception of the *implausible distractor* criteria. It appears that GPT-4 may have surpassed the crowd in this area due to its ability to quickly identify outliers in data sets (Su et al. 2024). This finding suggests that while human input remains crucial in the question quality evaluation process, automated methods could effectively handle specific criteria where their performance is comparable to that of humans. Implementing such a hybrid approach could reduce the workload for experts or crowdworkers; instead of assessing 19 IWF criteria per question, they might only need to evaluate 5, primarily confirming or refining the outputs from the automated evaluation. This could also lessen the demand for deep domain knowledge, as crowdworkers could focus on verifying the logic behind the AI's classifications, which provides a layer of human oversight to help mitigate the potential bias and errors introduced by the LLM (Ji et al. 2024).

Furthermore, in the more complex and detailed IWF rubric (19 criteria) compared to the SAQ evaluation (9 criteria), the performance was generally lower. The subjective nature of the *would you use it* criteria posed a particular challenge, especially for programmatic methods. It is difficult for both LLMs and humans to assess such a criterion effectively without substantial contextual information. Even less subjective criteria, like the *conciseness* of SAQs, showed low performance across all methods. This variability could be attributed to the diverse interpretations of *conciseness* among crowdworkers given their unique backgrounds.

Feasibility Assessing the feasibility of different methods for evaluating educational content, it becomes clear that neither experts nor crowdsourcing are cost-effective options. For instance, while Prolific achieves high results, the time and cost it took to set up the task to evaluate five questions in a typical online course is impractical.

Despite the costs and challenges, automated methods have shown promise, particularly as LLMs continue to advance. Yet, human computation still appears to be the most effective for evaluating MCQs and SAQs. Combining the two by integrating human insights with automated processes could optimize efficiency. For example, the use of GPT-4 could be integrated as part of a hybrid workflow, as it has demonstrated success by achieving perfect Micro F1 scores for several criteria and performing comparably to human evaluators in other aspects. This suggests a combined approach might alleviate some of the burdens on human evaluators by involving them only when necessary. While designing better questions from the start is ideal, there's a practical aspect to consider as well: many existing questions are already available in various banks and online courses (Costello, J. Holland, and Kirwan 2018a). Instead of creating new content from scratch, a more efficient approach could be to evaluate and improve existing questions. Crafting high-quality MCQs and SAQs is a skill that requires time and practice, and even LLMs occasionally produce flawed questions. Recognizing that no method is perfect, leveraging both automated and human resources could enhance the overall quality of educational assessments.

Limitations The inherent subjectivity associated with human ratings was addressed by employing verified and validated rubrics, yet some level of subjectivity inevitably remains. Additionally, the use of LLMs introduced potential biases related to their training data and algorithms. The task formulation itself, both for the crowdworkers and the LLMs, presented challenges, including the precise wording of rubric criteria and considerations regarding the native language of participants, which could affect their understanding and application of the rubrics. Particularly with LLMs, the various prompt wordings can drastically change the outputs as well, so consistent phrasing and temperature is crucial for reliable results.

Conclusion

This study explored the effectiveness and reliability of crowdsourced and programmatic methods for evaluating the quality of multiple-choice questions MCQs and SAQs across various educational domains. By leveraging the IWF rubric and a 9-item SAQ rubric, we systematically compared the performance of crowdworkers from MTurk and Prolific with three state-of-the-art LLMs: GPT-4, Gemini 1.5 Pro, and Claude 3 Opus. Our findings reveal that while crowdsourcing can harness wide-reaching human insights, LLMs offer a scalable alternative that approaches the reliability and accuracy of expert judgments. The application of standardized rubrics by both crowdworkers and LLMs highlighted the potential for a hybrid approach, combining the nuanced understanding of human reviewers with the efficiency and consistency of automated systems. This work highlights the trade-offs of each method and demonstrates the feasibility of integrating these approaches to improve the pedagogical value of assessments. As we move forward, refining these hybrid strategies could significantly enhance the way educational content is evaluated, ensuring both the scalability of the evaluation process and the quality of educational assessments.

References

Abdi, Solmaz, Hassan Khosravi, and Shazia Sadiq. 2020. "Modelling Learners in Crowdsourcing Educational Systems." Pp. 3–9 in *International Conference on Artificial Intelligence in Education*. Springer.

Addlesee, Angus, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. "Multi-Party Goal Tracking with LLMs: Comparing Pre-Training, Fine-Tuning, and Prompt Engineering." Pp. 229–41 in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Ahn, June, Ha Nguyen, Fabio Campos, and William Young. 2021. "Transforming Everyday Information into Practical Analytics with Crowdsourced Assessment Tasks." Pp. 66–76 in *LAK21: 11th International Learning Analytics and Knowledge Conference*.

Amidei, Jacopo, Paul Piwek, and Alistair Willis. 2018a. "Evaluation Methodologies in Automatic Question Generation 2013-2018." Pp. 307–17 in *Proceedings of the 11th International Conference on Natural Language Generation*.

Amidei, Jacopo, Paul Piwek, and Alistair Willis. 2018b. "Rethinking the Agreement in Human Evaluation Tasks." Pp. 3318–29 in *Proceedings of the 27th International Conference on Computational Linguistics*.

Arif, Taimoor, Sumit Asthana, and Kevyn Collins-Thompson. 2024. "Generation and Assessment of Multiple-Choice Questions from Video Transcripts Using Large Language Models." Pp. 530–34 in *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. Atlanta GA USA: ACM.

Assaly, Ibtihal R., and Oqlah M. Smadi. 2015. "Using Bloom's Taxonomy to Evaluate the Cognitive Levels of Master Class Textbook's Questions." *English Language Teaching* 8(5):100–110.

Azevedo, Jose Manuel, Ema P. Oliveira, and Patrícia Damas Beites. 2019. "Using Learning Analytics to Evaluate the Quality of Multiple-Choice Questions: A Perspective with Classical Test Theory and Item Response Theory." *The International Journal of Information and Learning Technology* 36(4):322–41.

Bates, Simon P., Ross K. Galloway, Jonathan Riise, and Danny Homer. 2014. "Assessing the Quality of a Student-Generated Question Repository." *Physical Review Special Topics-Physics Education Research* 10(2):020105.

Breakall, Jared, Christopher Randles, and Roy Tasker. 2019. "Development and Use of a Multiple-Choice Item Writing Flaws Evaluation Instrument in the Context of General Chemistry." *Chemistry Education Research and Practice* 20(2):369–82.

Butler, Andrew C. 2018. "Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning?" *Journal of Applied Research in Memory and Cognition* 7(3):323–31.

Cochran, Keith, Clayton Cohn, Nicole Hutchins, Gautam Biswas, and Peter Hastings. 2022. "Improving Automated Evaluation of Formative Assessments with Text Data Augmentation." Pp. 390–401 in *International Conference on Artificial Intelligence in Education*. Springer.

Costello, Eamon, Jane C. Holland, and Colette Kirwan. 2018. "Evaluation of MCQs from MOOCs for Common Item Writing Flaws." *BMC Research Notes* 11(1):849. doi: 10.1186/s13104-018-3959-4.

Costello, Eamon, Jane Holland, and Colette Kirwan. 2018. "The Future of Online Testing and Assessment: Question Quality in MOOCs." *International Journal of Educational Technology in Higher Education* 15(1):42. doi: 10.1186/s41239-018-0124-z.

Doroudi, Shayan, Ece Kamar, and Emma Brunskill. 2019. "Not Everyone Writes Good Examples but Good Examples Can Come from Anywhere." Pp. 12–21 in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7.

Douglas, Benjamin D., Patrick J. Ewell, and Markus Brauer. 2023. "Data Quality in Online Human-Subjects Research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA." *Plos One* 18(3):e0279720.

Fletcher, Sam, and Md Zahidul Islam. 2018. "Comparing Sets of Patterns with the Jaccard Index." *Australasian Journal of Information Systems* 22.

Gao, Mingqi, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. "LLM-Based NLG Evaluation: Current Status and Challenges."

Haladyna, Thomas M., Steven M. Downing, and Michael C. Rodriguez. 2002. "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment." *Applied Measurement in Education* 15(3):309–33.

Horbach, Andrea, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. "Linguistic Appropriateness and Pedagogic Usefulness of Reading Comprehension Questions." Pp. 1753–62 in *Proceedings of The 12th Language Resources and Evaluation Conference*.

Hwang, Kevin, Sai Challagundla, Maryam Alomair, Lujie Karen Chen, and Fow-Sen Choa. 2023. "Towards AI-Assisted Multiple Choice Question Generation and Quality Evaluation at Scale: Aligning with Bloom's Taxonomy." in *Workshop on Generative AI for Education*.

Ji, Jiaming, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. "Beavertails: Towards Improved Safety Alignment of Llm via a Human-Preference Dataset." *Advances in Neural Information Processing Systems* 36.

Jury, Breanna, Angela Lorusso, Juho Leinonen, Paul Denny, and Andrew Luxton-Reilly. 2024. “Evaluating LLM-Generated Worked Examples in an Introductory Programming Course.” Pp. 77–86 in *Proceedings of the 26th Australasian Computing Education Conference*. Sydney NSW Australia: ACM.

Kamalloo, Ehsan, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. “Evaluating Open-Domain Question Answering in the Era of Large Language Models.” Pp. 5591–5606 in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Kevian, Darioush, Usman Syed, Xingang Guo, Aaron Havens, Geir Dullerud, Peter Seiler, Lianhui Qin, and Bin Hu. 2024. “Capabilities of Large Language Models in Control Engineering: A Benchmark Study on GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra.”

Khosravi, Hassan, Kirsty Kitto, and Joseph Jay Williams. 2019. “RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities.” *Journal of Learning Analytics* 6(3):91–105.

Kobren, Ari, Thomas Logan, Siddarth Sampangi, and Andrew McCallum. 2014. “Domain Specific Knowledge Base Construction via Crowdsourcing.” in *Neural Information Processing Systems Workshop on Automated Knowledge Base Construction, AKBC, Montreal, Canada*.

Kremer, Ilan, Yishay Mansour, and Motty Perry. 2014. “Implementing the ‘Wisdom of the Crowd.’” *Journal of Political Economy* 122(5):988–1012.

Labutov, Igor, and Christoph Studer. 2017. “JAG: A Crowdsourcing Framework for Joint Assessment and Peer Grading.” in *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31.

van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. “Human Evaluation of Automatically Generated Text: Current Trends and Best Practice Guidelines.” *Computer Speech & Language* 67:101151. doi: 10.1016/j.csl.2020.101151.

Liu, Zhe, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2024. “Make LLM a Testing Expert: Bringing Human-like Interaction to Mobile GUI Testing via Functionality-Aware Decisions.” Pp. 1–13 in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. Lisbon Portugal: ACM.

Lu, Owen HT, Anna YQ Huang, Danny CL Tsai, and Stephen JH Yang. 2021. “Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students Learning Performance.” *Educational Technology & Society* 24(3):159–73.

Markel, Julia M., Steven G. Opferman, James A. Landay, and Chris Piech. 2023. “GPTeach: Interactive TA Training with GPT-Based Students.” Pp. 226–36 in *Proceedings of the Tenth ACM Conference on Learning @ Scale*. Copenhagen Denmark: ACM.

McHugh, Mary L. 2012. “Interrater Reliability: The Kappa Statistic.” *Biochemia Medica* 22(3):276–82.

Moon, Hyeongdon, Yoonseok Yang, Hangyeol Yu, Seunghyun Lee, Myeongho Jeong, Juneyoung Park, Jamin Shin, Minsam Kim, and Seungtaek Choi. 2022. “Evaluating the Knowledge Dependency of Questions.” Pp. 10512–26 in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Moore, Steven, Eamon Costello, Huy A. Nguyen, and John Stamper. 2024. “An Automatic Question Usability Evaluation Toolkit.” Pp. 31–46 in *Artificial Intelligence in Education*. Vol. 14830, *Lecture Notes in Computer Science*, edited by A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt. Cham: Springer Nature Switzerland.

Moore, Steven, Ellen Fang, Huy A. Nguyen, and John Stamper. 2023. “Crowdsourcing the Evaluation of Multiple-Choice Questions Using Item-Writing Flaws and Bloom’s Taxonomy.” Pp. 25–34 in *Proceedings of the Tenth ACM Conference on Learning @ Scale*. Copenhagen Denmark: ACM.

Moore, Steven, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. 2022. “Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3.” Pp. 243–57 in *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-^{TEL} 2022, Toulouse, France, September 12–16, 2022, Proceedings*. Springer.

Moore, Steven, Huy A. Nguyen, Tianying Chen, and John Stamper. 2023. “Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods.” Pp. 229–45 in *Responsive and Sustainable Educational Futures*. Vol. 14200, *Lecture Notes in Computer Science*, edited by O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, and T. Papathoma. Cham: Springer Nature Switzerland.

Mulla, Nikahat, and Prachi Gharpure. 2023. “Automatic Question Generation: A Review of Methodologies, Datasets, Evaluation Metrics, and Applications.” *Progress in Artificial Intelligence* 12(1):1–32. doi: 10.1007/s13748-023-00295-9.

Pate, Adam, and David J. Caldwell. 2014. “Effects of Multiple-Choice Item-Writing Guideline Utilization on Item and Student Performance.” *Currents in Pharmacy Teaching and Learning* 6(1):130–34.

Pi, Sainan, Xin An, Shuo Xu, and Jinghong Li. 2020. “A Comparative Study on Three Multi-Label Classification Tools.” Pp. 8–12 in *Proceedings of the 2020 3rd International Conference on Information Management and Management Science*. London United Kingdom: ACM.

Rush, Bonnie R., David C. Rankin, and Brad J. White. 2016. “The Impact of Item-Writing Flaws and Item Complexity on Examination Item Difficulty and Discrimination Value.” *BMC Medical Education* 16(1):1–10.

Singh, Anjali, Christopher Brooks, and Shayan Doroudi. 2022. “Learnersourcing in Theory and Practice: Synthesizing the Literature and Charting the Future.” Pp. 234–45 in *Proceedings of the Ninth ACM Conference on Learning @ Scale*. New York City NY USA: ACM.

Stahl, Maja, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. “Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation.” *arXiv Preprint arXiv:2404.15845*.

Steuer, Tim, Leonard Bongard, Jan Uhlig, and Gianluca Zimmer. 2021. “On the Linguistic and Pedagogical Quality of Automatic Question Generation via Neural Machine Translation.” Pp. 289–94 in *European Conference on Technology Enhanced Learning*. Springer.

Su, Jing, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024. “Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review.”

Sun, Lu, Yuhan Liu, Grace Joseph, Zhou Yu, Haiyi Zhu, and Steven P. Dow. 2022. “Comparing Experts and Novices for Ai Data Work: Insights on Allocating Human Intelligence to Design a Conversational Agent.” Pp. 195–206 in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 10.

Tarrant, Marie, Aimee Knierim, Sasha K. Hayes, and James Ware. 2006. “The Frequency of Item Writing Flaws in Multiple-Choice Questions Used in High Stakes Nursing Assessments.” *Nurse Education Today* 26(8):662–71.

Tarrant, Marie, and James Ware. 2008. “Impact of Item-Writing Flaws in Multiple-Choice Questions on Student Achievement in High-Stakes Nursing Assessments.” *Medical Education* 42(2):198–206.

Yavuz, Fatih, Özgür Çelik, and Gamze Yavaş Çelik. 2024. “Utilizing Large Language Models for EFL Essay Grading: An Examination of Reliability and Validity in Rubric-based Assessments.” *British Journal of Educational Technology* bjet.13494. doi: 10.1111/bjet.13494.

Yuan, Alvin, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P. Dow, and Bjorn Hartmann. 2016. “Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques.” Pp. 1005–17 in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. San Francisco California USA: ACM.

Zhang, Dell, Jun Wang, and Xiaoxue Zhao. 2015. “Estimating the Uncertainty of Average F1 Scores.” Pp. 317–20 in *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. Northampton Massachusetts USA: ACM.