# Developing an Automated Evaluation Tool for Multiple-Choice Questions

Steven Moore

*Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, Pennsylvania, 15213, United States*

**Abstract**

The use of multiple-choice questions (MCQs) in higher education has increased due to their efficiency, objective grading, ability to generate item-analysis data, and short response time. Recently, learnersourcing has emerged as a method to scale up MCQ creation by involving students in the question creation process. While previous research has shown that students can effectively generate high-quality questions, the evaluation of student-generated questions remains a challenge due to subjectivity in human evaluation. The Item-Writing Flaws (IWF) rubric provides a standardized way to evaluate MCQs, but its application has relied on experts, making it difficult to scale. With recent advances in natural language processing, it may be possible to automatically apply the IWF rubric to student-generated questions, providing real-time feedback to students during the question creation process. Additionally, the Bloom's Taxonomy level and knowledge components (KCs) of the questions can also be automatically mapped to these student-generated questions. The goal of this research is to develop a tool that provides automatic evaluation and feedback for student-generated questions in the learnersourcing process, resulting in the creation of higher quality questions in a more efficient manner. First, we will investigate methods for automatically assessing educational MCQs for their quality and pedagogical usefulness. Second, we propose an extensive two phase study to evaluate the effectiveness of our tool for generating and evaluating multiple-choice questions in higher education. This will be done by observing how students utilize the tool across different academic domains within online courseware.

**Keywords**

Learnersourcing, Question Generation, Automatic Question Evaluation

## 1. Introduction

Multiple-choice questions (MCQs) are a widely used form of assessment in higher education, both for formative and summative evaluations. MCQs are advantageous because of their efficiency to score, objective grading, ability to generate item-analysis data, and the shorter time required for students to respond [2]. In recent years, authoring educational MCQs has extended beyond instructors, and has been scaled up by leveraging students in the process of question creation [20]. This is known as learnersourcing, where students complete activities that produce new content that can then be leveraged by future students [20]. Previous research has demonstrated that despite having a range of expertise, students can effectively generate short-answer and multiple-choice questions that are of comparable quality to expert-generated ones [14]. Additionally, research supports the act of question generation as a beneficial learning activity, so the benefit is mutual. Learnersourcing efforts have also led to the creation of several systems that allow students to create and answer questions generated by their peers [6, 10].

These systems often rely on student evaluation of other student-generated questions, as a method to assess the quality and usefulness of the questions. A common challenge in human-evaluation of educational material is subjectivity, particularly when the expertise of the evaluator might not be inline with the question content. Evaluation of these student questions is needed though before other students work on them, as low quality questions can not only waste student time, but also be detrimental to their learning [18]. One common

Proceedings of the Doctoral Consortium of the 18th European Conference on Technology Enhanced Learning, 4th September 2023, Aveiro, Portugal

EMAIL: stevenmo@andrew.cmu.edu
ORCID: 0000-0002-5256-0339 (A. 1)

evaluation method is the Item-Writing Flaws (IWF) rubric that utilizes experts to evaluate questions. This rubric contains 19 different criteria and provides a standardized way to evaluate multiple-choice questions that accounts for their quality and pedagogical usefulness [1, 17]. The use of this rubric helps to avoid the common pitfalls of evaluating questions based on subjective features or evaluator opinion for what constitutes a questions' quality.

While the use of the IWF rubric is an effective way to evaluate multiple-choice questions, previous efforts have relied on experts to apply it, making it challenging to scale. However, with recent advances in the natural language processing domain, many of the criteria for the rubric can be applied using a series of rules implemented via most programming languages [13]. Therefore, it may be possible to automatically apply the criteria to student-generated questions at the time they are being generated, rather than evaluating them after the fact. In evaluating the student-generated questions as they are actively partaking in the authoring process by providing real time feedback, it can streamline the student generation of high quality multiple-choice questions and prevent the need for students to evaluate other student-generated questions. Leveraging recent advances, the Bloom's Taxonomy level of the questions could also be automatically mapped, along with the skills or knowledge components (KCs) required to solve the question [22]. After students use automatic feedback from the IWF rubric to create a high-quality question, they can be shown the potential Bloom's Taxonomy and KCs their question assesses and make any necessary changes or approvals to them.

Ultimately, if students' effort is going to be applied to a learning activity that involves them generating multiple-choice questions, then the process should be mindful of their time and learning during the process. Through receiving automated feedback as they generate the questions, student output should not only yield high quality questions, but ones that are mapped to a Bloom's Taxonomy label and set of skills or knowledge components required to answer the question. In doing so, when the questions are used by other students, the proper learning analytics can be leveraged to better monitor student learning. Towards this goal, we pursue these research questions:

1. Is it possible to automatically assess student-generated multiple-choice questions with the item-writing flaws rubric?
2. Can a tool be developed that provides automatic evaluation and feedback for student-generated questions in the learnersourcing process, alleviating the need for human evaluation?
3. To what extent does the use of the tool result in students creating higher quality questions in a more efficient manner compared to traditional methods?

# 2. Background

## 2.1. Automatic Question Evaluation

Educational MCQs generated by instructors, students, or automatically are all susceptible to flaws that impact their efficacy and quality [17]. One challenge in evaluating MCQs' quality lies in determining what criteria are sufficient to quantify a question as being high-quality and effective for use in an educational context. To overcome this subjectivity, different item response theory and statistical methods have been utilized to evaluate student-generated MCQs [9, 12]. However, these techniques require post-hoc analysis of student performance data, which can be detrimental to the learning process, because if the questions being used have not been first vetted for their quality, then they may be poorly constructed which can negatively impact students' performance and achievement [4]. To help overcome this, recent research has leveraged different methods for automatically evaluating questions.

The automatic evaluation of questions often utilizes metrics related to readability and explainability, including natural language processing (NLP) metrics like BLEU and METEOR [19]. However, these metrics were shown to not correlate with human evaluation and to not have pedagogical implications. Recent efforts towards automatic evaluation of educational questions have relied on large datasets of student responses, which are then used to train different classification models [15, 16]. Obtaining datasets across diverse subject areas poses challenges for these methods, which often rely on limited publicly available datasets consisting of basic reading comprehension or lower grade-level academic questions. These methods infrequently utilize questions from complex domains that go beyond the cognitive process of recall [12]. Additionally, the model architectures used in these methods often lack interpretability, due to their simplistic evaluation criteria or blackbox training methods.

## 2.2. Item-Writing Flaws Rubric

For evaluating the quality and pedagogical usefulness of educational multiple-choice questions, human evaluation remains as the benchmark [12]. While different rubrics have been employed for this evaluation process, the item-writing flaws (IWFs) rubric containing 19 criteria for assessing educational questions has been standardized and evaluated via previous research [1, 14, 17]. A previous study that utilized this 19-item IWFs rubric assessed the quality of over two thousand MCQs [21]. Utilizing two human evaluators, they determined that nearly half of the questions were deemed unacceptable for educational usage, due to having more than one IWF. In this case, the question difficulties may be skewed to be too easy or too hard, which in turn misleads students and related learning analytics [7]. While this rubric is effective at evaluating educational questions, the application of it often requires substantial human effort and is time-consuming, especially when evaluating large numbers of questions across multiple subject areas [8]. However, many of the rubric criteria can be automatically evaluated for questions, reducing much of this effort.

## 2.3. Labeling Bloom's Taxonomy and Knowledge Components

Multiple-choice questions are often used to assess lower levels of Bloom's Taxonomy, such as remember and understand. However, MCQs also have the potential to assess higher levels such as application and analysis when they are properly designed [8]. Assessing these higher levels of Bloom's Taxonomy is desirable, as the higher cognitive processes are associated with better student learning. Additionally, in order to solve a problem, a student must also possess a specific set of skills or knowledge components (KCs). KCs are formally defined as specific pieces of information necessary to solve a problem [11]. Recent research has demonstrated success in automatically classifying the Bloom's Taxonomy label of MCQs [22]. It has also shown promise in automatically suggesting KCs for questions. By combining these automated methods with previous work on learnersourcing KCs from students, it could lead to more expert-level results. Having the Bloom's Taxonomy label and the set of KCs for a question is beneficial in that it can fuel learning analytics systems to better measure student learning. These labels are also essential for many methods of measuring student learning and providing adaptivity, such as knowledge tracing [5].

## 3. Research Methods

The initial phase of this research will involve conducting a thorough literature review of current learnersourcing systems and automatic-question evaluation methods. This review will examine how the domain of the course that students are asked to generate questions in might impact the success. This will help to identify potential challenges that exist between different domains for question generation and inform how we can develop towards a domain agnostic evaluation method. We will also investigate the many different question evaluation criteria used in prior studies, focusing on criteria that are used in educational contexts and include the pedagogical applicability of the questions, such as the IWF rubric. Finally, the literature review will provide insights into how existing learnersourcing systems for question generation are used by students, which will help inform the design decisions of our tool. The review's findings will serve as a foundation for the multiple-choice question evaluation process that can be combined or expanded to develop a question authoring tool.

Once an initial version of the tool is developed, the sequential study has two phases: evaluation of existing question datasets and a user-study involving student utilization of the tool. For the evaluation using existing datasets, we will leverage educational multiple-choice questions from previous studies that have varying levels of question quality, such as from the PeerWise platform or the LearningQ dataset

[3, 6]. We will also leverage expert-generated questions from a variety of domains from courses on the Open Learning Initiative (OLI) platform, which can still contain potential flaws that our system should be able to identify. We will manually evaluate these questions using our defined question evaluation criteria. From there, we will run the automatic evaluation to determine which criteria we might need to improve upon. This will also help inform how our automatic evaluation is affected by different domains and question content, causing some criteria to be evaluated more successfully than others.

Following this, the user-study involving students will be deployed through the OLI platform, which hosts a plethora of open educational courses used at higher education institutions across the world. Through embedding our tool within OLI courses of different domains and with students of different knowledge levels, we can gain a diverse sample representative of more students. Students in the courses we select will opt in, as part of our IRB protocol, and in doing so they will be presented with an activity that has them generate a multiple-choice question as they work through certain parts of their respective course. We will utilize metrics collected from the platform, such as time on task, the quality of the student-generated questions, the amount of feedback they received from the tool, and which flaws students commonly encountered. These will help us determine not only if the tool helps students create high quality multiple-choice questions, but also if it benefits their learning and the types of students that are making these high quality questions. We also have previously collected student-generated questions from several existing OLI courses, where students created them without the use of a tool, receiving no feedback. The questions students create through our tool can further be compared to them, to help measure how much the tool helps or hinders this process for them.

## 4. Contributions

Learnersourcing continues to grow, as students have already authored over a million questions using tools that might not optimally support their learning or time as effectively as possible. This work will contribute an open source tool that can be used to help students author these questions. Additionally, the methods can be retroactively applied to questions from online courses and datasets to evaluate existing questions and indicate how they might be improved. Through the use of this tool, we will also contribute a dataset of student-generated questions from a variety of domains in higher education, which will cover more complex topics than existing educational multiple-choice question datasets. The results of this research will inform how we can better create questions across a variety of domains and make improvement to assessments that are actively being used by students in existing online courseware.

## 5. References

[1] Breakall, J., Randles, C. and Tasker, R. 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 (2019), 369–382. 2.

[2] Butler, A.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 (2018), 323–331.

[3] Chen, G., Yang, J., Hauff, C. and Houben, G.-J. 2018. LearningQ: a large-scale dataset for educational question generation. Twelfth International AAAI Conference on Web and Social Media (2018)

[4] Clifton, S.L. and Schriner, C.L. 2010. Assessing the quality of multiple-choice test items. Nurse educator. 35, 1 (2010), 12–16.

[5] Corbett, A.T. and Anderson, J.R. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction. 4, 4 (1994), 253–278.

[6] Denny, P., Hamer, J., Luxton-Reilly, A. and Purchase, H. 2008. PeerWise: students sharing their multiple choice questions. Proceedings of the fourth international workshop on computing education research (2008), 51–58.

[7] Downing, S.M. 2005. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement

examinations in medical education. Advances in health sciences education. 10, (2005), 133–143.

[8] Ji, T., Lyu, C., Jones, G., Zhou, L. and Graham, Y. 2022. QAScore—An Unsupervised Unreferenced Metric for the Question Generation Evaluation. Entropy. 24, 11 (2022)

[9] Khairani, A.Z. and Shamsuddin, H. 2016. Assessing item difficulty and discrimination indices of teacher-developed multiple-choice tests. Assessment for Learning Within and Beyond the Classroom: Taylor's 8th Teaching and Learning Conference (2015), 417–426.

[10] Khosravi, H., Kitto, K. and Williams, J.J. 2019. Ripple: a crowdsourced adaptive platform for recommendation of learning activities. arXiv preprint arXiv:1910.05522. (2019).

[11] Koedinger, K.R., Corbett, A.T. and Perfetti, C. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive science. 36, 5 (2012), 757–798.

[12] Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S. 2020. A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education. 30, (2020), 121–204.

[13] Moore, S., Nguyen, H.A., Bier, N., Domadia, T. and Stamper, J. 2022. Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. 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 (2022), 243–257.

[14] Moore, S., Nguyen, H.A. and Stamper, J. 2021. Examining the Effects of Student Participation and Performance on the Quality of Learnersourcing Multiple-Choice Questions. Proceedings of the Eighth ACM Conference on Learning @ Scale (2021)

[15] Ni, L., Bao, Q., Li, X., Qi, Q., Denny, P., Warren, J., Witbrock, M. and Liu, J. 2022. Deepqr: Neural-based quality ratings for learnersourced multiple-choice questions. Proceedings of the AAAI Conference on Artificial Intelligence (2022), 12826–12834.

[16] Ruseti, S., Dascalu, M., Johnson, A.M., Balyan, R., Kopp, K.J., McNamara, D.S., Crossley, S.A. and Trausan-Matu, S. 2018. Predicting question quality using recurrent neural networks. Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19 (2018), 491–502.

[17] Rush, B.R., Rankin, D.C. and White, B.J. 2016. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. BMC medical education. 16, 1 (2016), 1–10.

[18] Schurmeier, K.D., Atwood, C.H., Shepler, C.G. and Lautenschlager, G.J. 2010. Using item response theory to assess changes in student performance based on changes in question wording. Journal of chemical education. 87, 11 (2010), 1268–1272.

[19] Scialom, T. and Staiano, J. 2020. Ask to Learn: A Study on Curiosity-driven Question Generation. Proceedings of the 28th International Conference on Computational Linguistics (2020), 2224–2235.

[20] Singh, A., Brooks, C. and Doroudi, S. 2022. Learnersourcing in Theory and Practice: Synthesizing the Literature and Charting the Future. Proceedings of the Ninth ACM Conference on Learning@ Scale (2022), 234–245

[21] Tarrant, M., Knierim, A., Hayes, S.K. and Ware, J. 2006. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Education Today. 26, 8 (2006), 662–671.

[22] Wang, Z., Manning, K., Mallick, D.B. and Baraniuk, R.G. 2021. Towards blooms taxonomy classification without labels. Artificial Intelligence in Education: 22nd International Conference, AIED (2021), 433–445.