Quality Agreement on Learnersourced Multiple Choice Questions*

Richard Glassey[1,*,†], Olle Bälter[1,†]

[1]KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

Learnersourcing presents an efficient and economical pathway to producing more learning content, whilst engaging students more actively and deeply in their learning. However, it also presents new challenges to solve. Most of all, how best can we manage the variance in quality of content produced. In this work, we focus on the extent to which students and teachers agree on quality of learnersourced multiple choice questions. Students (n=30) were tasked with producing six questions over three weeks of an introductory programming course as part of their assessment. They also had to review 12 questions authored by their peers over the same period using a set of principles for good questions. After this period, four teaching staff involved with the course reviewed the student questions using the same process and principles. Inter-rater reliability statistics found overall positive agreement across principles, however this dropped to weaker agreement for principles aimed at more subjective and higher order concerns of question quality and quality of question feedback.

Keywords

Learnersourcing, Peerwise, Inter-rater Reliability

1. Introduction

Learnersourcing can be succinctly defined as crowdsourcing content from students in a learning context [1]. Different types of learnersourced content include: content annotation, resource recommendation, explanation of misconceptions, content creation, and aspects of evaluation, reflection and regulation [2]. In all cases, students are active in adding value to content and creating new content for the consumption of other students.

Creating learning content is a higher-order activity for students [3]. However, as with any student activity there will be a spectrum of engagement and quality. This creates a challenge of management as ideally students are exposed to the highest quality content, whilst lower quality content is filtered out of the system. Both PeerWise [4] and RiPPLE [5] have adopted content quality management strategies as platform features. However, one question that bubbles up is do students and teachers agree what quality is, given its subjective nature?

Here, we approach this question by studying the agreement gaps that emerge when students and teachers are asked to give their opinions about learnersourced MCQs. To give structure

Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY

[*] Corresponding author.

[†] These authors contributed equally.

glassey@kth.se (R. Glassey); ob1@kth.se (O. Bälter)

0000-0002-8996-0221 (R. Glassey); 0000-0001-5626-1187 (O. Bälter)

to these opinions, we use a set of principles for producing better MCQs that were presented during the production and review of questions [6]. Finally, we calculated inter-rater reliability statistics [7] to discover both the differences within student and teacher reviews and between student and teacher reviews.

2. Context of Study

In response to the 2015 European Refugee Crisis, academics at KTH Royal Institute of Technology, Stockholm, Sweden developed an intensive three month training to integrate newly arrived into the local IT workforce [8]. As this training was not attached to traditional course delivery, there was much more freedom to innovate and try novel pedagogical interventions [9]. In the first three weeks, when students were covering introductory programming, assessment was achieved by having students create multiple choice questions on the topics they were learning. The main motivation was to quickly generate a lot of MCQs that the students could then answer to increase their opportunities to practice. This was achieved and it was found that 50% of students answered 100 questions or more, without any demand from teachers to do so [6].

Early iterations of this training found that, whilst students could produce good MCQs there was a cold-start problem in MCQ quality - students gradually got better over time. Furthermore, students were good at writing questions and answering alternatives, however when it came to the explanation or feedback that accompanied the MCQ, students struggled to provide similar quality [6]. In response, we developed a set of 12 principles for writing good MCQs, reflecting the patterns of quality issue we detected in reviewing student MCQs [6], rather than more established guides developed for academics [10]. The principles are listed in table 1.

In the most recent iteration, to help students understand and apply the principles, students worked as a group to learn about writing good questions before they undertook the task individually. Principles were used to create a question and then review a question created by another group. In this way, students had a chance to discuss their interpretation of the principles and apply them both in creating and reviewing MCQs. For each week of the three week course, students were tasked with creating two questions and reviewing four questions. Once the course had ended, four teachers reviewed 96 questions independently using the principles and review process used by the students. For each question there were three student reviews and four teacher reviews.

3. Findings and Discussion

Table 1 shows the inter-rater reliability (IRR) results for both student and the teacher reviewers, according to each principle, sorted by the difference between students and teachers, and the final column shows the combined IRR for both students and teachers. IRR statistics provide the amount of agreement between independent reviewers when assessing the same things [7]; in our case, students and teachers rating MCQs according to principles. As there were more than two reviewers, Gwet's AC1/AC2 agreement coefficient amongst multiple raters was selected to calculate IRR [7].

Table 1

Principles of good multiple choice questions ranked by difference of student vs teacher inter-reviewer reliability (IRR). Students and teachers reviewers are shown separately, then their difference, and then finally when they are combined altogether. Green highlighting indicates strong agreement, grey indicates moderate agreement, and red indicates where there is weak agreement.

Principles of Good Multiple Choice Questions	Student IRR	Teacher IRR	Diff	All IRR
Only the feedback to the correct alternative reveals the answer	0.20	0.73	0.53	0.27
Question is aiming at higher order thinking	0.21	0.61	0.40	0.20
Feedback is unique and provided for each answer alternative	0.60	0.97	0.37	0.47
Question is formulated to ease readability	0.72	0.95	0.23	0.81
Question targets a misconception	0.57	0.69	0.12	0.52
All answer alternatives are formulated to ease readability	0.81	0.93	0.12	0.86
All answer alternatives are plausible and related to a misconception	0.59	0.69	0.09	0.46
Feedback is sufficient to understand why each alternative was incorrect or correct	0.48	0.54	0.06	0.25
Question is reasonable to solve without external systems	0.88	0.95	0.06	0.91
All feedback is correct	0.78	0.83	0.05	0.74
Question is from the course domain	0.94	0.98	0.04	0.94
Three or more answer alternatives are provided	0.97	1.00	0.03	0.97
	$\overline{x_1}$: 0.65	$\overline{x_2}$: 0.82		$\overline{x_3}$: 0.62

First, taking student and teacher IRR scores separately, teachers had a higher average agreement over all principles (0.82) versus students (0.65). Also, students had a greater distribution of scores (from 0.20 to 0.97) versus teachers (from 0.54 to 1.00). These findings can be partly explained by the teachers both generating the principles and also discussing their possible interpretations. Students on the other hand had much less opportunities to discuss the interpretation of the principles, other than within the scheduled group creation and review sessions, which only occurred twice in the three week course.

Second, looking at the largest differences in IRR scores by principle, "*Only the feedback to the correct alternative reveals the answer*" had a magnitude of 0.53. This suggests a difference in attitude with teachers seeing more value for MCQ feedback that helps correct student misconception without taking away the chance of a second attempt. Another large difference relating to the feedback (0.37), "*Feedback is unique and provided for each answer alternative*", continues this theme, whereas students did not have similar strong agreement. This is interesting as it is an objective measure - each answering alternative should have its own unique feedback and it only requires a cursory glance to confirm. Much like the smallest difference (0.03) for "*Three or more answer alternatives are provided*", where it is quite easy to count the number of answering alternatives. One limitation of PeerWise is that there is only a single text area for 'explanation' and this may have contributed to this anomaly.

Third, looking at the smallest differences that are not clearly objective or easy to determine (like "*Question is from the course domain*"), both principles: "*All answer alternatives are plausible and related to a misconception*" and "*Feedback is sufficient for you to understand why each alternative was incorrect or correct*" are interesting as these are quite challenging aspects of quality to agree upon, even for teachers who understand the intent of each principle. This suggests a limitation of the depth of quality one can hope to measure when rating learnersourced content, however the IRR results here are still promising with medium agreement between students (0.59 and 0.48) and between teachers (0.69 and 0.54).

Finally, when combining both student and teacher reviews together, creating a pool of seven reviewers, the most agreement can be found in perhaps the most objectively answerable principles. This is not surprising and acts as a nice control for agreement on the basics of MCQs. Of more concern is that where there is weak agreement, two principles are concerned with feedback ("*Feedback is sufficient for you to understand why each alternative was incorrect or correct*" and "*Only the feedback to the correct alternative reveals the answer*") and the other concerning if the question targets higher order thinking. This represents a challenge that warrants deeper investigation as the value of feedback is well known and accepted in effective education, but what if our different points of view on it never actually meet and the feedback that teachers feel is sufficient is not exactly (or even close to) what students need?

4. Conclusion

Learnersourcing generates more content, but comes with the challenge of how to determine quality. Part of solving the challenge is finding out where students and teachers agree (or not) about quality. Use of agreement statistics, such as inter-rater reliability, are a potentially useful metric, but as mentioned here, caution is advised as there are many to choose from and not all work as expected. The work presented here shows that there are areas we can find agreement, however we need to find better ways to solicit impressions of quality at deeper levels and then find ways to integrate and automate them within learnersourcing platforms.

References

[1] J. Kim, Learnersourcing: improving learning with collective learner activity, Ph.D. thesis, Massachusetts Institute of Technology, 2015.

[2] Y. Jiang, D. Schlagwein, B. Benatallah, A review on crowdsourcing for education: State of the art of literature and practice., PACIS (2018) 180.

[3] S. Abdi, H. Khosravi, S. Sadiq, G. Demartini, Evaluating the quality of learning resources: A learnersourcing approach, IEEE Transactions on Learning Technologies 14 (2021) 81–92.

[4] P. Denny, A. Luxton-Reilly, J. Hamer, The peerwise system of student contributed assessment questions, in: Proceedings of the tenth conference on Australasian computing education-Volume 78, 2008, pp. 69–74.

[5] H. Khosravi, G. Gyamfi, B. E. Hanna, J. Lodge, Fostering and supporting empirical research on evaluative judgement via a crowdsourced adaptive learning system, in: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, 2020, pp. 83–88.

[6] R. Glassey, O. Bälter, Put the students to work: Generating questions with constructive feedback, in: 2020 IEEE Frontiers in Education Conference (FIE), IEEE, 2020, pp. 1–8.

[7] K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology 61 (2008) 29–48.

[8] M. Wiggberg, E. Gobena, M. Kaulio, R. Glassey, O. Bälter, D. Hussain, R. Guanciale, P. Haller, Effective reskilling of foreign-born people at universities-the software development academy, IEEE Access 10 (2022) 24556–24565.

[9] R. Glassey, O. Bälter, Sustainable approaches for accelerated learning, Sustainability 13 (2021) 11994.

[10] T. M. Haladyna, S. M. Downing, M. C. Rodriguez, A review of multiple-choice item-writing guidelines for classroom assessment, Applied measurement in education 15 (2002) 309–333.