Does Student Crowdsourcing of Practice Questions and Animations Lead to Good Quality Materials?

Dr. Alex Daniel Edgcomb, University of California, Riverside

Alex Edgcomb finished his PhD in computer science at UC Riverside in 2014. Alex has continued working as a research specialist at UC Riverside with his PhD advisor, studying the efficacy of web-native content for STEM education. Alex also works with Zyante, a startup that develops interactive, web-native textbooks in STEM.

Joshua Sai Yuen, University of California, Riverside

Graduate student at University of California, Riverside

Prof. Frank Vahid, University of California, Riverside

Frank Vahid is a Professor of Computer Science and Engineering at the Univ. of California, Riverside. His research interests include embedded systems design, and engineering education. He is a co-founder of zyBooks.com.
Does Student Crowdsourcing of Practice Questions and Animations Lead to Good Quality Materials?

Alex Edgcomb, Joshua Yuen, and Frank Vahid
Dept. of Computer Science and Engineering, University of California, Riverside.
{aedgcomb, jyuen, vahid}@cs.ucr.edu
Also with zyBooks.com

Abstract

Web-native textbooks use practice questions and animations to improve student performance to help learn and visualize concepts. Creating practice questions and animations is time intensive. This paper investigates whether having students create and/or rate practice questions and animations—student crowdsourcing—can lead to good quality items.

For animations, we conducted experiments involving 587 participants from a basic computing technology course. Students were asked to create animations of a topic, and a professor and fellow students rated those animations. Some students showed the ability to create quality animations: 2 of 19 completed animations were rated a "5" on a 1 to 5 scale (5 is best) by a professor / professional animation author. Furthermore, some students showed the ability to effectively rate student-created animations; the top 10% of student ratings was strongly correlated with the professor ratings with R-value = 0.88 (p-value < 0.001).

For questions, we conducted experiments involving 25 participants from an introductory embedded programming course. Students were asked to create and rate practice questions for various embedded programming topics. Some students could effectively rate questions: the average of the top 20% of student ratings was strongly correlated with the professor rating with R-value = 0.82 (p-value = 0.02). However, students did not show the ability to create good question; no student's question was professor-rated above a 4. The common problem seen was an inability to write correct and precise English.

Keywords: Crowdsourcing, rating, authoring, student-made content, web-native content, interactive content, programming, computer science, digital learning, digitally-enhanced education, online learning.

Introduction

Textbooks are giving way to online interactive material[15][20][23]. Interactive material benefits student learning with immediate feedback and higher levels of engagement[8], which may be a better match to a modern student's learning style.

However, interactive material is hard to make, whether authoring a new textbook or converting an existing textbook. One possibility is to have students help create interactive
material. In this paper, we investigate whether students can create good-quality interactive materials.

We expect that most students will not make good-quality interactive material, but we hope that perhaps some students can, and that such good materials could percolate to the top of a list via peer reviews. Then, a professional editor can curate the top interactive materials. Such student crowdsourcing enables large amounts of material to be created at scale, proportional to the number of contributing students. Numerous education groups\(^3\)[5][9][11][12][18] have developed similar processes for rating videos, applications, practice questions, interactive exercises, and images.

In this paper, we consider two types of interactive material: animations and practice questions. An animation illustrates a concept with moving objects. Typically, an animation replaces the pictures in a textbook with multiple moving substeps. Animations can convey many concepts in science, technology, engineering, and mathematics (STEM). However, animation creation is time intensive, typically 1-2 hours each. We investigate whether students can make good-quality animations.

A practice question provides quick feedback to a student on whether the student understands a concept, and may also teach new concepts. Students have shown a preference to practice questions over reading\(^8\).

First, we investigate whether students can reliably rate the quality of practice questions, and whether students can create good-quality practice questions. Then, we investigate whether students can create good-quality educational animations, and whether students can reliably rate the quality of educational animations. We found that some students could reliably rate practice questions and animations, especially better raters averaged together. Also, some students could create good-quality animations, but almost no students could create good-quality practice questions.

**Background**

Numerous online groups have crowdsourced the rating of digital education resources\(^3\)[5][9][11][12][18][21]. Texas Computer Educators Association (TCEA)\(^18\) used a 1-4 (4 is best) evaluation rubric that studies 6 main areas of a digital application: curriculum connection, authenticity, feedback, differentiation, user friendliness, and student motivation. Common Sense Media (CSM)\(^5\) used a 0-5 (5 is best) scale to evaluate engagement levels, and learning approaches of the applications for users ages 2-17. Applications that received a 0 rating were considered not suitable for users in the 2-17 age range or for learning purposes. The evaluators involved for TCEA and CSM were teachers and educators. Gooru Learning\(^11\) used a user based 1-5 (5 is best) rating approach where a user submits a review for an educational material. Wikipedia\(^21\) has replaced all traditional encyclopedias with a crowdsourced approach involving different tiers of users. Wikipedia has over 22 million Wikipedians (Wikipedia contributors) that regularly edit, and contribute to new or pre-existing resources. Each Wikipedian has access to edit or create his or her own local resources, which other users, who have a
history of reliable edits, may then approve as global resources. Wikipedia allows the use of images, text, and links to other sources. Khan Academy[12] has volunteers that help develop practice problems and micro-lectures.

Academic groups have investigated the importance of crowdsourcing of educational material[1][6][19]. Weld[19] identified companies, such as Khan Academy, that use crowdsourced development of content to grow and mature the content. Anderson[1] describes higher education as two groups of knowledge audiences: knowledge-seekers and knowledge-providers. When mentors are unavailable, knowledge-seekers turn to knowledge-providers for instruction and aide. Content generated by knowledge-providers may be crowdsourced and vetted by knowledge-seekers in a game-like rating scheme to ensure reliability. Similarly, Corneli[6] proposed the idea of a role system for learners, where learners are all seeking different knowledge and understanding of specific topics. Corneli suggests that most learner roles may benefit from a crowdsourced model but does not provide experimental evidence.

Academic groups used forms of collaborative crowdsourced learning to study the effectiveness of student’s assessment and peer instruction[7][10][13][14][16]. de Alfaro[7] created a crowdsourced grading tool, CrowdGrader, that allows students to grade and review their peer's homework submissions. CrowdGrader was found to actively involve students in grading other's assignments. O'Neil[14] found that with the use of collaborative class lecture notes, students created high-quality lecture notes when provided with a lecture skeleton layout. Notes created by the students also reflected how students were understanding the content in the course. Kumar[13] studied the effectiveness of an online tutor that provided questions to a student and then graded the student's answer with feedback. Students showed a 30-60% improvement from pre-quiz to post-quizzes. Ghauth[10] developed a system for recommending e-learning content based on learner's performance through a set of several different exams. The recommended e-learning content that was chosen was based on learner's ratings of the content as well as their performance. Ghauth concluded that with this recommendation system and rating system, 12.16% of users showed a significant improvement in understanding topics. Ssemugabi[16] compared two usability evaluation methods of web-based learning applications between 4 expert evaluators and 61 learners. Ssemugabi discovered that the expert evaluators identified 91% more problems with the web-based e-learning application than the learners. In this case, learners were insufficient for discovering problems with a web-based e-learning application.

Several companies have applied crowdsourcing to problem solving[2][17][21][22]. Amazon’s Mechanical Turk[2] is a crowdsourcing Internet marketplace for hiring (or being hired to perform) human intelligence tasks, such as identifying parked cars in a picture taken on a rainy night. Bian[4] discussed companies, such as Wikipedia[21] and Yahoo![22], that introduced Community Question Answering (CQA), which include the ability for users to rate other user’s content for quality and reliability. Bian added a supervisor component to the CQA concept and discovered higher quality answers to the asked questions.
Practice question crowdsourcing

Practice questions are intended to help learn written material of a textbook’s section. Our practice question crowdsourcing experiments tested whether students could reliably rate the quality of practice questions, and whether students could create good-quality practice questions. In the first experiment, student ratings were compared to professor-made ratings on the quality of various practice questions. In the second experiment, students wrote practice questions that were then rated on quality by a professor. Each experiment contained the same group of participants.

Participants

The participants were undergraduate students at University of California at Riverside enrolled in Introduction to Embedded Systems (IES) during the 2014 second summer session. Traditionally, students in IES are juniors or seniors majored in computer science, computer engineering, or electrical engineering. 25 students participated in the two experiments. 4 participants were excluded from the first experiment due to not completing practice question rating, leaving 21 participants for the first experiment. 2 participants were excluded from the second experiment for not completing the practice question creation, leaving 23 participants in the second experiment. Both experiments were assigned as homework in IES, each worth a full homework assignment, thus self-selection bias was mitigated. The participants were blind to the conditions and specific purpose of the experiment. The experiment was approved by University of California at Riverside's IRB (Institutional Review Board).

Design

In the first experiment, participants rated the practice questions on a 5-point scale, from 1 to 5 (5 is best). The participants were shown 7 professor-created practice questions and 7 student-created practice questions. The participants did not know which practice questions were created by a professor and which by a student. The professor was a professional author who had made thousands of questions for various online learning content. In the second experiment, the participants created practice questions that were rated by the professor on a 5-point scale from 1 to 5 (5 is best).

Materials

Both experiments consisted of a series of webpages that students worked through at their own pace as homework assignments. The first experiment consisted of a question-quality rating tutorial to train students how to rate questions, followed by a question-rating page. The second experiment consisted of a question submission form, question-writing tutorial, and then a question-writing tool.

The question-quality rating tutorial contained a 5-item guideline, shown in Figure 1. The items were: word choice clarity, concise/precise wording, proper grammar, practice question explains answer, and practice question teaches relevant concept. For each item
in the guideline that was satisfied, the participants were instructed to award one point to the practice question. The participant could reference the tutorial at any time during the experiment.

The question-rating page was located below the question-quality tutorial. The question-rating page contained 14 practice questions. 7 practice questions were student-made and 7 were professor-made. IES students in Spring 2014 created the student-made practice questions. The order of the practice questions were randomized.

**Figure 1.** Question rating guideline from the question-rating tutorial. Participants were instructed to award a question one point for each satisfied item in the guideline.

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word choice</td>
<td>Does not use words like I, you, we, your, my. Avoids vague words/phrases like it, they, that means, and this means.</td>
</tr>
<tr>
<td>Concise/precise</td>
<td>Concise, precise question and explanation. Uses minimal words. Ensure short answer questions have only one possible correct answer.</td>
</tr>
<tr>
<td>Grammar</td>
<td>Grammatically correct.</td>
</tr>
<tr>
<td>Explains</td>
<td>Explanation should expound on why an answer is right or wrong.</td>
</tr>
<tr>
<td>Teaches</td>
<td>Question teaches a challenging and relevant topic.</td>
</tr>
</tbody>
</table>

Each practice question contained a question, an answer, and an explanation. As shown in Figure 2, three practice question formats were used: true-false (Figure 2(a)), multiple choice (Figure 2(b)), and short answer (Figure 2(c)). For true-false, the explanation given was the same whether true or false was clicked. For multiple choice, each clicked choice provided a different explanation. For short answer, a user enters an answer into a text input and has the options of clicking the Show answer or Check button. Clicking the Show answer button twice reveals the answer and explanation. Clicking the Check button when the text input has the right answer reveals the answer and explanation. Clicking the Check button when the text input has the wrong answer reveals a hint. The hint is the same for any wrong text input.

**Figure 2.** (a) True-false, (b) multiple choice, and (c) short answer question. (a) has the incorrect answer selected. (b) and (c) have the correct answer.
The question submission form contained instructions how to complete the second homework assignment, i.e., the second experiment. The participant was asked to enter his or her email (to receive credit for the homework assignment), to review the question-writing tutorial, then to use the question-writing tool to write 5 practice questions. Each practice question was to be associated with a particular section in the course book. The first 3 practice questions were to be in the following order: short answer, multiple choice, then true-false. The latter 3 practice questions could be any practice question format that the participant chose. For each practice question, a blank text area was provided for copying in practice question written in the question-writing tool. At the end of the question submission form was a Submit button. If the Submit button was clicked and the participant’s email and each practice question text area had content, then participant was notified that the homework was completed. Otherwise, the participant was prompted to complete the empty text area.

The question-writing tutorial contained the guideline shown in Figure 1, as well as, a brief introduction to writing each practice question format: true-false, multiple choice, and short answer. Each introduction contained a sentence and a picture shown an example practice question.

The question-writing tool was a form, with a different template for each practice question format. The true-false template included a text area for the question, the option to select true or false as the expected answer, editable labels for “True” and “False”, and a text area for the explanation. The multiple-choice template included a text area for the question and a button to add more options. Each option included an answer text area and explanation text area, and a remove button. The short answer format included a separate text area for the question, answer, hint, and explanation.

*Procedure*

The instructor of IES assigned both experiments as homework in the 2014 summer session at the University of California at Riverside. Participants had 7 days to work on the first experiment, and 4 days to work on the second experiment. The participants could complete the experiments at their own pace and work on the experiments any time or place the participant preferred.

*Results and discussion*

The practice question crowdsourcing experiments tested whether students could reliably rate the quality of practice questions, and whether students could create good-quality practice questions.

The correlation between the participant ratings and professor rating is the level of consistency between the ratings. The ideal correlation is 1.0, meaning the participant and professor would agree on the quality of an practice question. Each participant’s ratings of a practice question was correlated (to produce an R-value) with the professor’s rating of
the same practice question. 6 out of the 21 (or 29%) R-values exceeded 0.5 (the highest R-value was 0.77), and 3 out of those 6 had a p-value < 0.05, indicating that some participants individually had a medium-strength correlation with the professor.

The participant’s ratings were averaged per practice question. This average was correlated with the professor ratings for the same practice question resulting in an R-value of 0.46 and p-value = 0.20.

**Figure 3.** Top-20% student ratings compared to professor ratings. Each point is one practice question. The regression line shows the rating consistency.

![Graph showing correlation between top-20% student ratings and professor ratings.](image)

More importantly, the top 4 (or 20%) of participant’s ratings (top defined as having the highest individual R-value) were averaged per practice question. The top-20% ratings were correlated with the professor ratings, resulting in an R-value of 0.82 with p-value = 0.02. The top-20% R-value is close to the ideal of 1.0 and considered a strong correlation. Figure 3 shows the relationship graphically. The drawn line is the regression line, indicating the relationship.

Toward whether students can create good-quality practice questions, a professor rated 90 practice questions written by participants. 0% were rated as 5 out of 5, 4.4% were rated as 4 out of 5, 25.6% were rated as 3, 41.1% were rated as 2, and 28.9% were rated as 1. The participant-made practice questions commonly contained the following mistakes: practice question was conceptually incorrect; practice question mixed concepts in a way that the answer cannot be determined; improper English usage leading to wrong meaning; poorly designed or awkward practice question; trivial practice question leading to no learning; incorrect English and/or wording.

Figure 4 shows two participant-made practice questions. Figure 4(a) is a better question, rated at 4 out of 5, and could have been improved with more a more concise question wording and being more precise about what qualifies as a major difference, or avoiding the issue of major difference all together. Figure 4(b) is a bad question, rated at 1 out of 5, and is conceptually confused about the relationship between a loop and a state machine. Also, Figure 4(b)’s usage of “must” leads to a wrong meaning.
Animation crowdsourcing

An animation is intended to visually describe a particular concept with animated shapes and text. The animation crowdsourcing experiments tested whether students could create good-quality educational animations, and whether students could reliably rate the quality of educational animations. The experiments consisted of three parts, each with a unique group of participants:

1. Creation of animations
2. Rating of all on-topic created animations
3. Rating of top-20 on-topic created animations

Participants

The participants were undergraduate students at the University of California at Riverside enrolled in a basic computing course (BCC) during the last week of the 2013 Fall quarter. BCC introduces basic computing applications such as Microsoft Office and web applications, intended for non-computing and non-engineering majors and having no prerequisites. Traditionally, 98% percent of BCC students are non-computing and non-engineering majors. There were 587 participants in total, split across 3 groups. Each participant was in just 1 group. 136 participants created an animation. 334 participants rated from all on-topic created animations. 117 participants rated from top-20 on-topic created animations. The participants were blind to the conditions and specific purpose of the experiments. The participants were participating as part of their computing applications coursework, thus reducing self-selection bias. The experiments were approved by University of California at Riverside's IRB (Institutional Review Board).

Design

The first group of participants created one animation each of a specified topic. A researcher from this work and two research assistants evaluated the 136 animations created by students. An animation was excluded for not being on-topic (typically meaning the student just created a fun or nonsense animation and did not try to animate the specified topic), of which 29 animations were on-topic.

Figure 4. Participant-made practice questions. (a) Professor rated 4 out of 5, and (b) professor rated 1 out of 5. Both practice questions have the correct answer selected, with the explanation shown in green.
The second group of participants were asked to rate 20 animations, randomly-selected per participant, from the 29 on-topic animations. The animation ratings were on a 1 to 5 scale, where 5 was best.

The third group of participants were asked to rate all 20 of the top-20 on-topic animations. 1 professional-made animation was combined with the top-19 rated animations from the 29 on-top animations to create the top-20 on-topic animations. The professional-made animation was created by two of the researchers from this work and feedback from two researcher assistants and an author of a web-native textbook.

Materials

The experiments were conducted in the normal lab classrooms during normal lab hours on the last week of the Fall 2013 quarter. The lab had approximately 40 desktop computers with identically configured Linux and Mozilla’s Firefox installed. Each student was randomly assigned a unique identifier that was associated with each animation rating and each animation creation. The unique identifier was stored in temporary browser cache.

The experiments consisted of 6 webpages: a landing page, an animation creation page, an animation creation manual, an animation creation follow-up survey, an animation rating page, and an animation rating follow-up survey. Each participant started an experiment at the landing page. Then, the participant either: went to the animation creation manual followed by the animation creation page followed by the creation follow-up survey, or went to the animation rating page followed by the follow-up survey.

The landing page was titled: Animation Study. The landing page included that, “the purpose of this study is to evaluate the efficacy of an animation tool and crowd sourcing of animations.” Also, the landing page included: “Your responses are completely anonymous. You will be graded on participation only.” The participants were asked to “Please work alone.”

Figure 5. Explanation from the animation creation manual that includes some text and a 16-second GIF. There were 10 explanations in the manual.
The animation creation manual introduces the concept of an animation for education with an example. The manual then links to the animation creation page, followed by 10 explanations of how to use the animation tool. The explanations include how to create, delete, and modify an object in the animation, as well as, use the move and fade instructions. Each explanation consisted of a sentence or two, and either a picture, GIF, or video, such as Figure 5. Lastly, the manual includes another link to the animation creation page.

The animation creation page started with the animation goal of explaining swap: “Spend 20 mins building an animation that illustrates the concept of swapping the location of two items between your hands by temporarily setting down one item. For example, you have a book in one hand and a phone in the other. To swap, you place the phone on a table, move the book to the other hand, then pick up the phone.” Next, the creation page suggested using different colored objects, labeling the objects, and using instructions to move and fade objects. Lastly, the creation page included the animation tool, a feedback area, and a submit button. If a participant clicked the submit button and had been on the creation page for 20 minutes and had at least 6 objects in the animation, then the participant’s animation would submit and the webpage would direct to the animation creation follow-up survey. Otherwise, if the participant clicked the submit button, then a pop-up would appear saying that the animation complexity was insufficient, and suggest the participant consult the manual for more ideas.

**Figure 6.** Good-quality participant-made animation teaching the concept of swap. The phone has been set down temporarily on the table. The book is about to switch hands, then the phone will be picked up again.

The animation creation follow-up survey asked for suggestions to improve the animation creation tool, the student provide an email if interested in pursuing animation creation, and any other comments. The animation rating page contained a randomly-selected animation, such as the animation shown in Figure 6, that a participant could watch and re-watch. Below the animation was the prompt: “The above is intended to show the concept of swapping two objects while using a temporary location. For example, you have a book in one hand and a phone in the other. To swap, you place the phone on a table, move the book to the other hand, then pick up the phone.” Next, the question “How
well is the concept taught?” was written above a rating scale with 1-to-5 stars, followed by an open-ended feedback text area and a submit button. If the participant clicked submit without giving a rating, then the pop-up “Rating is a required field” would appear. Otherwise, another randomly-selected animation would load that had not been rated by the participant previously. After 20 animations were rated, then the webpage was automatically redirected to the animation rating follow-up survey. The animation rating follow-up survey asked the student to provide an email if interested in pursuing animation creation, and to provide any other comments.

Procedure

The participants attended his or her normal lab section, same room and time, during the last week of the Fall 2013 quarter. The lab assignment was the experiment. The participants were allotted the full 3-hour lab section, but most spent 30 to 45 minutes to complete the assignment. The participants were instructed to work alone. The normal teaching assistants monitored the participants. The teaching assistants were instructed not to help create or rate animations for the participants. Instead, the teaching assistants would suggest the student re-read the manual and lab instructions. Participants that completed the assignment were allowed to remain in the lab to work on other coursework, but not communicate with participants still working on the lab assignment.

Results and discussion

The animation crowdsourcing experiments tested whether students could create good-quality educational animations, and whether students could reliably rate the quality of educational animations.

Toward whether students could create good-quality education animations, a professor rated the top-20 on-topic animations, which included 1 professional-made animation unknown to the professor. Only the professional-made and two participant-made animations were rated with a 5 out of 5 by the professor. One participant-made animation is illustrated in Figure 6. The animations rated with 5 out of 5 included the following good-qualities: labeled the objects being swapped and the temporary position; either labeled or visually indicated the positions at which the objects begin and end; objects being swapped were concrete things, e.g. phone and book, as opposed to abstract things, e.g. “Object 1” “Object 2”; concisely demonstrated the concept; used meaningful colors to visually distinguish the objects.

Toward whether participants could reliably rate the quality of animations, 117 participants rated at least 1 animation, 36 of which were disqualified for not completing all ratings. Therefore, 81 participants were considered. The correlation between the participant ratings and professor rating is the level of consistency between the ratings. The ideal correlation is 1.0, meaning the participant and professor would agree on the quality of an animation. Each participant’s ratings of an animation was correlated (to produce an R-value) with the professor’s rating of the same animation. 10 out of the 81 (or 12%) R-values exceeded 0.5 (the highest R-value was 0.74; p-value = 0.002), and 9
out of those 10 had a p-value < 0.05, indicating that some participants individually had a medium-strength correlation with the professor.

The participant’s ratings were averaged per animation. This average was correlated with the professor ratings for the same animation resulting in an R-value of 0.74 and p-value = 0.002. This R-value is the same as the highest individual participant R-value, meaning the entire group of participants was as effective at rating as the single best individual participant.

**Figure 7.** Top-10% student ratings compared to professor ratings. Each point is 1 animation. The ratings were strongly correlated, as seen by the consistency with the regression line.

Further, the top 8 (or 10%) of participant’s ratings (top defined as having the highest individual R-value) were averaged per animation. The top-10% ratings were correlated with the professor ratings, yielding an R-value of 0.88 with p-value < 0.001. The top-10% R-value is close to the ideal of 1.0 and considered a strong correlation. Figure 7 shows the relationship graphically. The drawn line is the regression line, indicating the relationship.

Participants on average tended to rate with more 3s, 4s, and 5s, and less 1s and 2s, than the professor ratings. For example, the lowest average participant rating for an animation was 3.5 and highest was 4.4, whereas the professor used a fairly even distribution of the 1-5 ratings (three 1s, five 2s, four 3s, five 4s, and three 5s), resulting in differences in the average and standard deviation of average participant and professor ratings. Perhaps students were concerned about giving lower ratings to fellow students, or did not have enough confidence. Nonetheless, the student ratings strongly correlated with the professor ratings because the relative quality ratings were consistent.

**Conclusion**

Some students can reliably rate practice questions and animations. For practice questions, the average of the top 20% of student ratings strongly correlated with professor ratings with R-value = 0.82 (p-value = 0.02). For animations, the average of the top 10% of student ratings correlated with professor ratings with R-value of 0.88 (p-value < 0.001).
Some students can create good-quality animations, but students were not able to make good-quality practice questions. 2 of the 19 student-made animations rated by the professor achieved a 5 out of 5 rating. Good-quality animations are more about artistic abilities and creativity, whereas practice questions require precise thought and English usage. 4 of the 90 student-made practice questions achieved a 4 out of 5 rating from a professor; none achieved 5 out of 5. Creating a good-quality practice question requires more conceptual understanding and precise language usage than creating a good-quality animation, which may explain why students can make good-quality animations but not practice questions.

Future work includes investigating instructor-created materials using a crowdsourced evaluation technique as done above.

Acknowledgement

This material is based upon work supported by the National Science Foundation under SBIR Grant Nos. 1315094 and 1430537.

References