# Evaluating Crowdsourcing and Topic Modeling in Generating Knowledge Components from Explanations

Steven Moore(✉), Huy A. Nguyen, and John Stamper

Carnegie Mellon University, Pittsburgh, PA 15213, USA
StevenJamesMoore@gmail.com

**Abstract.** Associating assessment items with hypothesized knowledge components (KCs) enables us to gain fine-grained data on students' performance within an ed-tech system. However, creating this association is a time consuming process and requires substantial instructor effort. In this study, we present the results of crowdsourcing valuable insights into the underlying concepts of problems in mathematics and English writing, as a first step in leveraging the crowd to expedite the task of generating KCs. We presented crowdworkers with two problems in each domain and asked them to provide three explanations about why one problem is more challenging than the other. These explanations were then independently analyzed through (1) a series of qualitative coding methods and (2) several topic modeling techniques, to compare how they might assist in extracting KCs and other insights from the participant contributions. Results of our qualitative coding showed that crowdworkers were able to generate KCs that approximately matched those generated by domain experts. At the same time, the topic models' outputs were evaluated against both the domain expert generated KCs and the results of the previous coding to determine effectiveness. Ultimately we found that while the topic modeling was not up to parity with the qualitative coding methods, it did assist in identifying useful clusters of explanations. This work demonstrates a method to leverage both the crowd's knowledge and topic modeling to assist in the process of generating KCs for assessment items.

**Keywords:** Knowledge component · Knowledge component modeling · Crowdsourcing · Topic modeling · Intelligent tutoring systems

## 1 Introduction

The combination of data-driven knowledge tracing methods and cognitive-based modeling has greatly enhanced the effectiveness of a wide range of educational technologies, such as intelligent tutoring systems and other online courseware. In particular, these systems often employ knowledge component modeling, which treats student knowledge as a set of interrelated KCs, where each KC is "an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks" [14]. Operationally, a KC model is defined as a mapping between each question item and a

© Springer Nature Switzerland AG 2020
I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12163, pp. 398–410, 2020.
https://doi.org/10.1007/978-3-030-52237-7_32

hypothesized set of associated KCs that represent the skills or knowledge needed to solve that item. This mapping is intended to capture the student's underlying cognitive process and is vital to many core functionalities of educational software, enabling features such as adaptive feedback and hints [22].

While machine learning methodologies have been developed to assist in the automatic identification of new KCs, prior research has shown that human judgment remains critical in the interpretation of the improved model and acquisition of actionable insights [19, 24]. An emerging area that has the potential to provide the human resources needed for scaling KC modeling is crowdsourcing. Naturally, the challenge with this approach is that the population of crowdworkers is highly varied in their education level and domain knowledge proficiency. Therefore, as a first step towards examining and promoting the feasibility of crowdsourced KC modeling, we studied how crowdworkers can provide insights into different word problems that might suggest areas of improvements and generating KCs for the questions. We took these insights via explanations, coded them and ran them through two topic models to analyze how they might be utilized for the task. Our research questions are as follows:

*RQ1*: Are the explanations provided by crowdworkers indicative of any KCs that the problems require?
*RQ2:* How effective is topic modeling compared to qualitative coding in identifying explanations indicative of KCs?
*RQ3*: Do the explanations provide insights into how the presented assessment items may be improved?

## 2 Related Work

KC models are typically developed by domain experts through Cognitive Task Analysis methods [29], which lead to effective instructional designs but require substantial human efforts. Fully automated methods can potentially discover models with better performance than human-generated ones (in terms of statistical metrics such as AIC, BIC and cross validation score), but they suffer from a lack of interpretability [31]. Other efforts of automatic cognitive model discovery make use of student data, such as the Q-matrix algorithm [2]. On the other hand, [13] showed that a refined KC model that results from both human judgment and computational metrics can help students reach mastery in 26% less time. More generally, as pointed out in [18] the inclusion of human factors in the KC modeling process can be advantageous, leading to lessons that can be implemented in follow-up studies.

Recently, crowdsourcing has become increasingly popular for content development and refinement in the education domain [21, 27]. The process of crowdsourcing data from learners, or learnersourcing, has been used to identify which parts of lecture videos are confusing [12], and to describe the key instructional steps and subgoals of how-to videos [11]. In particular, [33] explored a crowdsourcing-based strategy towards personalized learning in which learners were asked to author explanations on how to solve statistics problems. The explanations generated by learners were found to be comparable in quality to explanations produced by expert instructors.

As the fields of natural language processing and text mining continue to advance, they are being increasingly leveraged by education to help automate arduous tasks [6]. Previous work has looked at using different machine learning models [25, 26] and utilizing a search engine [10] to tag educational content with KCs. Recent efforts have utilized topic modeling on a set of math problems from an intelligent tutoring system to assist in the labeling of KCs [30]. While their initial model had promising results, there was an issue of human interpretability for the topics it produced, that may be relieved by different models [17]. Much of the work in this space is focused towards predicting KCs for content, after being trained on similarly KC tagged problem. Few studies have tried to leverage text mining techniques to generate KCs for content, with no training or prediction modeling involved.

## 3 Methods

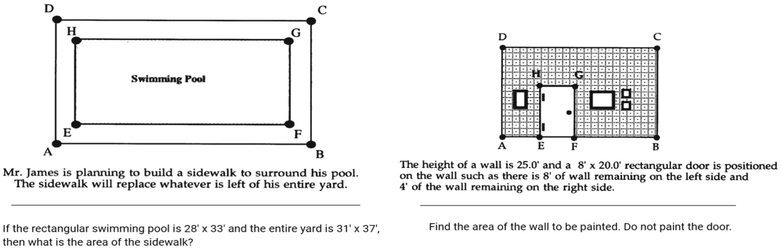
Our study consists of two experiments with the same procedure, but involve different domain knowledge. The first domain is mathematics, with a focus on the area of shapes; the second is English writing, with a focus on prose style involving agents and clause topics. In both domains, we deployed an experiment using Amazon's Mechanical Turk (AMT). Forty crowd workers on AMT, known as "turkers," completed the math experiment, and thirty turkers completed the writing experiment, for a total of 70 participants. In each domain, the tasks took roughly five minutes. Participants were compensated $0.75 upon completion, providing a mean hourly wage of $9.

The main task of the experiment presented participants with two word problems positioned side by side, labeled Question 1 and Question 2. In the math experiment, both problems involve finding the area of two different structures. In the writing experiment, both problems involve identifying the agents and actions of two different sentences. Participants were truthfully told that past students were tested on these problems and that the collected data indicates Question 2 is more difficult than Question 1. They were then asked to provide three explanations on why this is the case. The specific question prompt stated: *"Data shows that from the two questions displayed above, students have more difficulty answering Question 2 than Question 1. Please list three explanations on why Question 2 might be more difficult than Question 1"*.

### 3.1 Math and Writing Experiments

The two mathematics word problems used for the explanation task can be seen in Fig. 1. These problems come from a previous study of a geometry cognitive tutor [32], where the data indicates that students struggle more with the problem involving painting the wall (the right side of Fig. 1). Both problems are tagged with the same three KCs by the domain experts that created the problems, so they assess the same content. These KCs are: Compose-by-addition, Subtract, and Rectangle-area.

Both problems used in the writing experiment come from an online prose style course for freshman and sophomore undergraduates (Fig. 2). Similar to the math problems, student data collected from the online course indicates students struggle more with one problem over the other. The KCs were generated by domain experts and are: Id-clause-topic, Discourse-level-topic, Subject-position, and Verb-form.



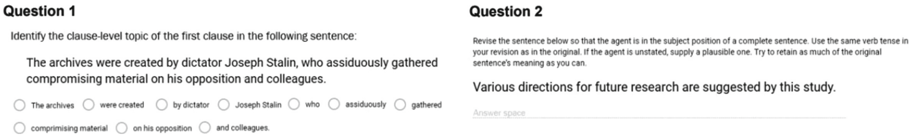
Mr. James is planning to build a sidewalk to surround his pool.
The sidewalk will replace whatever is left of his entire yard.

If the rectangular swimming pool is 28' x 33' and the entire yard is 31' x 37', then what is the area of the sidewalk?

The height of a wall is 25.0' and a 8' x 20.0' rectangular door is positioned on the wall such as there is 8' of wall remaining on the left side and 4' of the wall remaining on the right side.

Find the area of the wall to be painted. Do not paint the door.

**Fig. 1.** The two word problems for which participants provided three explanations in the math experiment, with the one on the right being more difficult.



**Question 1**

Identify the clause-level topic of the first clause in the following sentence:

The archives were created by dictator Joseph Stalin, who assiduously gathered compromising material on his opposition and colleagues.

○ The archives ○ were created ○ by dictator ○ Joseph Stalin ○ who ○ assiduously ○ gathered
○ comprimising material ○ on his opposition ○ and colleagues.

**Question 2**

Revise the sentence below so that the agent is in the subject position of a complete sentence. Use the same verb tense in your revision as in the original. If the agent is unstated, supply a plausible one. Try to retain as much of the original sentence's meaning as you can.

Various directions for future research are suggested by this study.

Answer space

**Fig. 2.** The two problems for which participants provided three explanations in the writing experiment, with the one on the right being more difficult.

### 3.2 Categorization of Explanations

We collected three explanations from each of the 40 participants in the math experiment, for a total of 120, and three explanations from each of the 30 participants in the writing experiment, for a total of 90. Overall there were 210 explanations, where each explanation is defined as the full text provided by a participant into the answer space. These mostly consisted of sentence fragments or full sentences, but there were several that had multiple sentences. Such explanations were still treated as a single unit, to which the best fitting code was applied [9].

Using data collected from a brief pilot study, two researchers followed the process in [7] to develop a codebook from the explanations in the math experiment, and a separate codebook for the writing experiment. This involved assigning the participant explanations to a set of codes based on their interpreted meaning. These codebooks were iteratively refined until agreement on the codes was achieved. Two research assistants then applied the codebook to the pilot data and discussed discrepancies, seeking clarity for any codes they were unfamiliar with. Table 1 shows the finalized version of the codebook applied to the collected math and writing explanation data. The codebook was then applied to the full dataset from each domain by the two research assistants. Next, we measured the code agreement via Inter-Rater Reliability (IRR). The coders achieved a Cohen's kappa $\kappa = 0.813$ for the math experiment and $\kappa = 0.839$ for the writing experiment, which indicates a high level of agreement [15].

### 3.3 Topic Modeling Explanations

Topic models estimate latent topics in a document from word occurrence frequencies, based on the assumption that certain words will appear depending on potential topics in

**Table 1.** Coding dictionary for the math and writing experiment responses.

| Code | Definition | Example explanation |
|---|---|---|
| *Math experiment* | | |
| Calculation | Mentions the computational aspects involved in the problem, e.g., subtraction or use of area | "Because they don't know how to calculate the area" |
| Clarity-Shape | Relates to the understanding of the depicted shape | "It may be less clear which part should be calculated because of shading" |
| Clarity-Text | Relates to the understanding of the text | "Wording is kinda confusing" |
| Complexity | Claiming that one problem is more complicated than the other, without further clarification | "Problem two is more complicated than problem one" |
| Composite | Addresses an embedded shape used in the problem | "The picture itself shows other objects such as windows and this might throw off the student" |
| Content | General remarks about the problem content that are not captured by other content subcategories | "The numbers displayed have decimal points" |
| Meta | A mention of general skills needed to solve any type of word problem, such as focusing, reading, and attention | "It takes more time to read in problem 2 so students are more prone to getting discouraged" |
| N/A | Does not provide any sensible explanation | "340" |
| Shape-Layout | Mentions the visual element of the word problem's shapes | "It is more difficult based on the shapes presented in question two" |
| Step-Num | Indicates one problem requires a certain number of steps/more steps | "There are more steps to complete in problem 2" |
| Value-Num | Indicates one problem has more variables/values to work with | "It has more variables" |
| *Writing experiment* | | |
| Answer # | Relating to the number of answer choices present in the question | "In option one there is only one right answer" |
| Complexity | Discusses the general difficulty/complexity | "More complex knowledge needed" |
| Content | Touches on the content of the question | "They have to revise it instead of just saying what is wrong" |
| Meta | Describing a skill required by similar problems, at a more meta level | "It is hard to write" |

(*continued*)

**Table 1.** (*continued*)

| Code | Definition | Example explanation |
|---|---|---|
| N/A | Not applicable or relevant | "Poor communication with suppliers" |
| Prework | Discusses the prior knowledge or prework that might be required to answer | "The second isn't explained in the coursework" |
| Question-type | Addresses the question's type (MCQ or free response) in the explanation | "Written answer instead of multiple choice" |
| Question-text | Mentions the question's text in some capacity, e.g., longer/confusing | "Sentence 2 is more vague" |
| Rules | Mentions the rules a student would need to know to solve the problem | "Problem one only requires an understanding of grammar" |
| Technical | Mentions a specific technical term that might be required to answer | "In problem two, the subject is not in the beginning of the sentence" |

the text. We used two topic modeling techniques, Latent Dirichlet Analysis (LDA [5]) and Non-negative Matrix Factorization (NMF [16]), to further analyze the explanations. LDA maps all documents, in this case the explanations, to a set number of topics in a way such that the words in each document are captured by the topics [1]. NMF uses linear algebra for topic modeling by identifying the latent structure in data, the explanations, represented as a non-negative matrix [20]. The explanation text was lemmatized and stop words were removed, using a common NLP library in Python [4]. No further text processing was performed on the explanation data before running them through the models, as we wanted results without fine-tuning any parameters or heavily processing the data. The results of the topic models were then evaluated against the researcher-generated codes, categorizations, and the expert generated KCs for the problems, in order to gauge their effectiveness for this task.

## 4 Results

*RQ1:* Are the explanations provided by crowdworkers indicative of any KCs that the problems require? From the coded explanations in the math and writing experiments, we constructed a set of themes, shown in Table 2, formed by grouping several of the related codes within each experiment together [28]. In the math experiment the first three themes, Greater Quantity, Shapes Present, and Domain Knowledge, all comprise explanations which address features of the given problems and are indicative of a KC required to solve the problem. Explanations that are grouped into these three themes can be translated into KCs that fit the problem and are indicative of the underlying skill(s) required to solve it. However, the only explanations that suggested a KC that matched any of the expert ones (Compose-by-addition, Subtract, and Rectangle-area) came from the *Calculation* code. The fourth theme, Clarity/Confusion, pertains to the problem's question text or visuals being unclear and hard to decipher. This theme contains explanations that relate to what makes the problems particularly difficult outside of the knowledge required to solve it; from these explanations, one could also derive ways to improve the assessment, such as making the question text more explicit or clarifying the depicted image. The fifth theme, Irrelevant, holds the remaining explanations – those that do not address the problem in a meaningful way, i.e., they are too general or abstract.

**Table 2.** Themes for the math (above) and writing (below) experiments created from the coded data and if the theme is akin to a KC or an area of problem improvement.

| Theme (# of explanations) | Codes | KC | Improvement |
|---|---|---|---|
| Greater quantity 27 | Step-num, value-num | ✔ | |
| Shapes present 30 | Shape-layout, composite | ✔ | |
| Domain knowledge 33 | Content, calculation | ✔ | |
| Clarity/confusion 15 | Clarity-text, clarity-shape | | ✔ |
| Irrelevant 15 | Complexity, meta, N/A | | |
| Process to solve 13 | Rules, content | ✔ | |
| Domain knowledge 07 | Prework, technical | ✔ | |
| Question specific attributes 42 | Question-text, question-type, answer-num | | ✔ |
| Irrelevant 28 | Complexity, meta, N/A | | |

In the writing experiment the first two themes, Process to Solve and Domain Knowledge, are indicative of KCs that were required to solve the problems. The only explanations that matched any of the expert generated KCs (Id-clause-topic, Discourse-level-topic, Subject-position, and Verb-form) for the problems came from the *Rules* and *Technical* codes. The third theme, Question Specific Attributes, discusses the relative level of difficulty between problems, due to one being multiple-choice and the other being free-response, or the question text differences between the two. This theme relates explanations that address ways to improve the assessment, such as simplifying the answer choices. Finally, the Irrelevant theme again consists of explanations that are not meaningful or overly general.

*RQ2:* How effective is topic modeling compared to qualitative coding in identifying explanations indicative of KCs? The 10 topics identified by both the LDA and NMF models, along with the five most common words associated with them, are presented in Table 3. From the math experiment data, both the LDA and NMF models had comparable results to one another. They share the same set of topic interpretations and an equally low number of N/A topics. While certain topics in both models are attributed to KC codes, it would be challenging to discern the explicit KC just from the terms. The three primary themes across the ten topics from each model are calculation of area, the visual nature of the shapes in the problems' figures, and how one problem is generally more complicated than the other. We expected some of the expert-generated KCs for the math problems (Compose-by-addition, Subtract, & Rectangle-area) to be identifiable in the topics. Surprisingly *'subtract'* was not a top five term for any topic nor was *'area'* a term alongside *'rectangle'* for any topics.

Similar to the math topics, both the LDA and NMF models produced comparable results for the writing experiment, with slightly different terms used for the topics between the two. The predominant topic in both models is related to the question type, which is appropriate as it was a dominating category from the qualitative coding. Interestingly, there are not as many topics involving *Complexity* or *N/A*, both irrelevant codes that attribute little to no meaning. The majority of the topics focus on the high-level features of the questions, such as the wording or type. Topic 9 from the LDA model and topic 7 from the NMF one include vocabulary used in two of the expert generated KCs (Id-clause-topic, Discourse-level-topic, Subject-position, and Verb-form). However, these topics and the others are not interpretable enough to discern such KCs explicitly from the terms.

*RQ3:* Do the explanations provide insights into how the presented assessment items may be improved? In addition to some of the explanations being indicative of a KC, such as ones that fall into the *Calculation* or *Technical* codes, many of the other explanations suggested complications with the word problems. In the math experiment, 15 of the 120 total explanations (12.5%) fall into the *Clarity/Confusion* theme from Table 2. Additionally, only 15 of the 120 (12.5%) were deemed *Irrelevant* to the problems, meaning that in general the majority of the explanations were either suggestive of an improvement that could be made or a KC required to solve them. The writing experiment had a greater number of explanations, 42 out of 90 (46.67%), that fell into the *Question Specific Attributes* theme in Table 2. Only 28 of the 90 (31.11%) explanations in this experiment were deemed *Irrelevant* to the problems.

## 5 Discussion and Implications

Firstly, we wanted to see if the provided explanations could be used to generate fitting KCs for the problems. We found that many of the provided explanations did address the underlying concepts required to solve a problem, more so in the math domain than the writing domain. For example, explanations from the math experiment in the *Greater Quantity* theme often discuss how one problem required the area calculation of more shapes than the other. Solving a problem that involves the area of multiple shapes instead of just a single one has been identified as a knowledge component for similar problems from a previous study [32]. This type of difficulty may be overlooked due to expert blindspot, as the explicit steps taken to solve a problem can get grouped together when it becomes second nature [23]. Eliciting the crowd for explanations such as these can help bring in a diverse level of knowledge, ranging from novice to expert, that can help to make this KC explicit.

From the writing experiment, the *Process to Solve* theme consists of the most KC indicative explanations. These often discuss a step required to solve one of the problems, which was usually at the granularity that would make it a fitting KC. Unfortunately the explanations contributed by participants that were indicative of KCs were relatively rare, making up only 20 of 90 (22.22%) of the total explanations from the writing data,

**Table 3.** Top 5 terms from 10 topics identified by the LDA and NMF topic models

| Topic # | LDA terms | LDA topic interpretation | NMF terms | NMF topic interpretation |
|---|---|---|---|---|
| *Math experiment* | | | | |
| 1 | Figure, question, hard, shape, confusing | Clarity-shape | Problem, longer, figure, steps, lines | Step-Num |
| 2 | Problem, complicated, 1, complex, 2 | Complexity | Area, windows, given, figure, door | Calculation |
| 3 | Step, calculation, need, require, work | Step-num | Confusing, wording, question, painted, wall | Complexity |
| 4 | Consider, answer, visually, complicated, simple | Shape-layout | Shapes, deal, irregular, question, rectangles | Shape-Layout |
| 5 | Width, 223, calculate, problem, attention | Calculation | Numbers, deal, size, work, need | N/A |
| 6 | Area, complicated, window, 143, 2 | Clarity-shape | Complicated, calculation, somewhat, problem, involves | Complexity |
| 7 | Confusing, know, abstract, somewhat, term | Complexity | Simple, question, involves, consider, shape | Complexity |
| 8 | Accommodate, time, difficult, shading, shape | Clarity-shape | Harder, visually, figure, shape, make | Clarity-Shape |
| 9 | Instruction, measurement, equal, forward, straight | N/A | Areas, account, figure, need, just | Calculation |
| 10 | Detail, variable, 340, long, contain | N/A | Difficult, calculate, solve, door, width | Calculation |
| *Writing experiment* | | | | |
| 1 | Answer, prework, specific, pick, confine | Prework | Choice, multiple, problem, allows, simple | Question-type |

(*continued*)

**Table 3.** (*continued*)

| Topic # | LDA terms | LDA topic interpretation | NMF terms | NMF topic interpretation |
|---|---|---|---|---|
| 2 | Multiple, choice, 1, problem, thinking | Question-type | Sentence, meaning, needs, subject, problem | Rules |
| 3 | Sentence, vague, problem, option, right | Question-text | Problem, requires, understanding, rules, thinking | Meta |
| 4 | Long, response, 1, free, variable | Question-type | Answer, free, easier, pick, right | Question-type |
| 5 | Know, comment, paraphrase, range, contain | Rules | People, writing, hard, write, questions | Meta |
| 6 | People, write, simplified, question, multiple | N/A | Comments, written, eliminate, like, level | N/A |
| 7 | Need, complex, written, knowledge, number | Complexity | Know, subject, verb, tense, agent | Technical |
| 8 | Comment, problem, choice, multiple, complex | Question-type | Answers, correct, just, questions, incorrect | Question-type |
| 9 | Comment, clause, look, agent, suggest | Technical | Clause, concept, agent, ended, like | Technical |
| 10 | Concept, rewrite, choose, sentence, end | Content | Complex, concept, written, ended, like | Complexity |

compared to 73 of 120 (60.83%) from the math domain. We attribute this difference between domains due to the knowledge required for them, as the math problems were from a middle school class and the writing questions from a college-level writing course.

The two topic models were only able to identify a few topics, each relating to *Calculation*, that fit into a code indicative of a KC that matched one an expert generated. While the terms for the topics can be gleaned for words that suggest a KC such as “area” or “window”, they still lack interpretability and a direct translation into a KC. This is also true of the two models’ results in the writing domain, which identified several topics relating to the *Rule* and *Technical* codes. Without further interpretation, the terms suggest some vocabulary used in the problems, but they are insufficient to derive an actionable KC without further human processing.

Secondly, we wanted to see if the explanations provided insights into how the assessment items might be improved. Both experiments had one theme directly related to improving the surface level features of the problems, such as the question text or images. For instance, in the math experiment, the theme *Clarity/Confusion* addresses the confusion caused by the visual elements of the problems. The included images for the questions are a key aspect to the assessment and beneficial to problem solving, but may be misinterpreted in a way the content creators may not have intended [8]. Correcting the images can allow for better assessments; based on the explanations we received, a student may answer incorrectly purely based on the poor image design.

Across both domains, the 10 topics identified by each model are mostly comprised of those that indicate areas of problem improvement. While the models performed poorly at generating KCs from the explanations, many of the topics and terms were indicative of student struggle due to confusion with the text or image of the problems. In total, 12.5% of the explanations in math and 31.11% in writing were considered irrelevant to the task and presented problems. Even with limited instruction and the varying backgrounds, participants were able to provide insights into the problems that could be used for baseline KC generation or identifying areas of assessment refinement.

## 6 Conclusion and Future Work

In this study, we gathered explanations for the relative difficulty between two mathematics questions and between two English writing questions from crowdworkers. We found that crowdworkers were able to generate valuable explanations that were indicative of a KC required to solve the problems or a suggestion for how to make the problems clearer. Understandably, they were able to provide better explanations in the easier domain of middle school math than in an undergraduate English writing domain. However, in both experiments, a majority of the explanations either pertained to identifying a KC or area of improvement, rather than being irrelevant. The LDA and NMF models created topics akin to the researcher generated codes, although the interpretability of these topics based solely on the terms is limited in usefulness. Nevertheless, the categories from the coding and topic models ultimately assisted in clustering explanations that were either indicative of a KC or an aspect of the problem that could be improved.

For future work, we plan to integrate this process in a learner-sourced context, where participants (i.e., students) potentially have more commitment and domain knowledge that could be leveraged [27]. This would enable us to properly train them to provide such explanations throughout the course, rather than completing the task once with only a brief instruction like the crowdworkers did in this study. Ultimately, we envision a workflow in which students submit explanations for why certain problems are difficult; these explanations are then peer reviewed and presented to the teachers (or relevant parties) to help them identify potential KCs and improve the assessment items. This procedure is analogous to the find-fix-verify pattern in crowdsourcing, which has been shown to be effective [2]. However, before reaching this point, the interpretability of the models will need to be improved or another technique should be utilized. This study demonstrates a first step in developing such a workflow, providing initial insights into how crowdsourced explanations might be leveraged for KC generation and assessment content refinement.

**Acknowledgements.** The research reported here was supported in part by a training grant from the Institute of Education Sciences (R305B150008). Opinions expressed do not represent the views of the U.S. Department of Education.

## References

1. AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C.: Topic significance ranking of LDA generative models. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS (LNAI), vol. 5781, pp. 67–82. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04180-8_22
2. Barnes, T.: The Q-matrix method: mining student response data for knowledge. In: American Association for Artificial Intelligence 2005 Educational Data Mining Workshop, pp. 1–8 (2005)
3. Bernstein, M.S., et al.: Soylent: a word processor with a crowd inside. In: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, pp. 313–322 (2010)
4. Bird, S., et al.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc. (2009)
5. Blei, D.M., et al.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Brack, A., et al.: Domain-independent extraction of scientific concepts from research articles. arXiv Prepr. arXiv:200103067 (2020)
7. DeCuir-Gunby, J.T., et al.: Developing and using a codebook for the analysis of interview data: an example from a professional development research project. Field Methods **23**(2), 136–155 (2011)
8. Edens, K., Potter, E.: How students "unpack" the structure of a word problem: graphic representations and problem solving. Sch. Sci. Math. **108**(5), 184–196 (2008)
9. Elliott, V.F.: Thinking about the coding process in qualitative data analysis. Qual. Rep. **23**, 11 (2018)
10. Karlovčec, M., Córdova-Sánchez, M., Pardos, Z.A.: Knowledge component suggestion for untagged content in an intelligent tutoring system. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 195–200. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30950-2_25
11. Kim, J., et al.: Learnersourcing subgoal labeling to support learning from how-to videos. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 685–690. ACM (2013)
12. Kim, J., et al.: Understanding in-video dropouts and interaction peaks in online lecture videos. In: Proceedings of the First ACM Conference on Learning@ Scale Conference, pp. 31–40. ACM (2014)
13. Koedinger, K.R., et al.: Automated student model improvement. 5th Int. Educ. Data Min. Soc. (2012)
14. Koedinger, K.R., et al.: The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. Cogn. Sci. **36**(5), 757–798 (2012)
15. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
16. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562 (2001)
17. Lee, T.Y., et al.: The human touch: how non-expert users perceive, interpret, and fix topic models. Int. J. Hum.-Comput. Stud. **105**, 28–42 (2017)

18. Liu, R., et al.: Interpreting model discovery and testing generalization to a new dataset. In: Educational Data Mining 2014. Citeseer (2014)
19. Liu, R., Koedinger, K.R.: Closing the loop: automated data-driven cognitive model discoveries lead to improved instruction and learning gains. J. Educ. Data Min. **9**(1), 25–41 (2017)
20. Luo, M., et al.: Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
21. Moore, S., et al.: Crowdsourcing explanations for improving assessment content and identifying knowledge components. In: Proceedings of the 14th International Conference of the Learning Sciences (2020)
22. Moore, S., Stamper, J.: Decision support for an adversarial game environment using automatic hint generation. In: Coy, A., Hayashi, Y., Chang, M. (eds.) ITS 2019. LNCS, vol. 11528, pp. 82–88. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22244-4_11
23. Nathan, M.J., et al.: Expert blind spot: when content knowledge eclipses pedagogical content knowledge. In: Proceedings of the Third International Conference on Cognitive Science (2001)
24. Nguyen, H., et al.: Using knowledge component modeling to increase domain understanding in a digital learning game. In: Proceedings of the 12th International Conference on Educational Data Mining, pp. 139–148 (2019)
25. Pardos, Z.A., Dadu, A.: Imputing KCs with representations of problem content and context. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 148–155 (2017)
26. Patikorn, T., Deisadze, D., Grande, L., Yu, Z., Heffernan, N.: Generalizability of methods for imputing mathematical skills needed to solve problems from texts. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11625, pp. 396–405. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23204-7_33
27. Paulin, D., Haythornthwaite, C.: Crowdsourcing the curriculum: redefining e-learning practices through peer-generated approaches. Inf. Soc. **32**(2), 130–142 (2016)
28. Saldana, J.: An introduction to codes and coding. Coding Man. Qual. Res. **3**, 1–31 (2009)
29. Schraagen, J.M., et al.: Cognitive Task Analysis. Psychology Press (2000)
30. Slater, S., et al.: Using correlational topic modeling for automated topic identification in intelligent tutoring systems. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. pp. 393–397 (2017)
31. Stamper, J., et al.: A comparison of model selection metrics in datashop. In: Educational Data Mining 2013 (2013)
32. Stamper, J.C., Koedinger, K.R.: Human-machine student model discovery and improvement using datashop. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNAI), vol. 6738, pp. 353–360. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21869-9_46
33. Williams, J.J., et al.: Axis: generating explanations at scale with learnersourcing and machine learning. In: 2016 Proceedings of the Third ACM Conference on Learning@ Scale, pp. 379–388. ACM (2016)