So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

# Utilizing Crowdsourcing and Topic Modeling to Generate Knowledge Components for Math and Writing Problems

**Steven MOORE[a*], Huy NGUYEN[b] & John STAMPER[c]**
[abc]*Human-Computer Interaction Institute, Carnegie Mellon University, United States*
*StevenJamesMoore@gmail.com

**Abstract:** Combining assessment items with their hypothesized knowledge components (KCs) is critical in acquiring fine-grained data on student performance as they work in an ed-tech system. However, creating this association is an arduous process and requires substantial instructor effort. In this study, we present the results of crowdsourcing KCs for problems in the domain of mathematics and English writing, as a first step in leveraging the crowd to expedite this task. We presented crowdworkers with a problem in each domain and asked them to provide three underlying skills required to solve it. These inputs were then analyzed through two topic modeling techniques, to compare how they might cluster around potential KCs. Results of the models' output were evaluated against KCs generated by domain experts to determine their usability. Ultimately, we found that half of the crowdsourced KCs matched expert-generated KCs in each problem. This work demonstrates a method to leverage the crowd's collective knowledge and topic modeling methods to facilitate the process of generating KCs for assessment items, which can be integrated in future learnersourced environments.

**Keywords:** Knowledge Component, Crowdsourcing, Topic Modeling, Text Mining, Knowledge Tracing, Intelligent Tutoring Systems

## 1. Introduction

Many educational technologies, such as intelligent tutoring systems and online courseware, utilize knowledge component modeling to support their adaptivity. This 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" (Koedinger, Corbett, and Perfetti, 2012). Operationally, a KC model is defined as a mapping between each question item and a 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 an intelligent educational software, enabling features such as adaptive feedback and hints (Moore and Stamper, 2019).

The construction of such a mapping is typically carried out by learning science practitioners, such as subject matter experts, cognitive scientists and learning engineers, who inspect the materials and assign one or more KCs to each question. Once student data has been collected, the initial mapping can then be improved as poorly associated KCs come to light (Corbett and Anderson, 1994). As the next step, instructional designers are often leveraged to revise the KC model; however, this is often a time-consuming task, making continuous iteration challenging. 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 (Liu and Koedinger, 2017; Nguyen et al., 2019).

An emerging area that has the potential to provide the human resources needed for scaling KC modeling is crowdsourcing. Over the last decade, crowdsourcing has become a prominent model of distributed problem-solving, especially on tasks that are doable by humans, but difficult for computers (Yuen, King, and Leung, 2011). It has also seen wide usage in academic research as a source of training data for machine learning, support for integrated workflows, and assistance to domain experts (Kobren et al., 2014). However, to our knowledge, there has been minimal prior work that employs crowdsourcing in KC model construction and refinement (Moore et al., 2020a).

Therefore, as a first step towards examining and promoting the feasibility of crowdsourced KC modeling, we studied how crowdworkers can provide the underlying KCs for an assessment activity. Using a crowdsourcing platform, we gathered participants with no background in pedagogical training and varying levels of math and English writing expertise. We then asked them to provide three KCs needed to solve a given problem in the domains of math and English writing. We took these contributions and input them to two topic models to analyze how they might be clustered in useful ways, indicative of any KCs. Our research questions are as follows:

**RQ1**: How do crowdsourced knowledge components compare to expert-generated ones in each problem?

**RQ2**: How might we leverage topic modeling to generate groupings of participant contributes indicative of expert-generated knowledge components?

From these questions, our goal is to see whether it is possible to employ crowdsourcing to generate a baseline KC model that is both interpretable and able to be translated into a more learnersourced context.

## 2. Related Work

KC models are typically developed by domain experts through Cognitive Task Analysis methods (Schraagen, Chipman, and Shalin, 2000), which lead to effective instructional designs but require substantial human efforts. In recent decades, this method has been greatly enhanced by data-driven techniques. For instance, (Corbett and Anderson, 1994) showed that identifying the "blips" or "peaks" in the KCs' learning curves can help uncover new KCs that were not accounted for in the initial model. Other works have formalized the concept of KC mapping through q-matrix or skill matrix (Koedinger et al., 2004). Under this representation, several data mining techniques can be applied to assist in the refinement of KC models, for example by extracting relevant KCs from data and merging/splitting KCs to yield a better fit (Koedinger, McLaughlin, and Stamper, 2012). It should be noted that while these fully automated methods can potentially discover models with better performance (in terms of statistical metrics such as AIC, BIC and cross validation score), they suffer from a lack of interpretability (Stamper, Koedinger, and Mclaughlin, 2013). On the other hand, (Koedinger et al., 2013) 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 (Liu, McLaughlin, and Koedinger, 2014), the inclusion of human factors in the KC modeling process can be advantageous, leading to actionable lessons that can be implemented in follow-up studies. Our work belongs to the same line of research, in which we try to leverage human insights, albeit from crowdworkers instead of traditional teachers and educators.

Traditionally, crowdsourcing can be carried out in either a marketplace, e.g. Mechanical Turk, a game environment, e.g. FoldIt (Cooper et al., 2010), or an online community, e.g. Stack Overflow. In this study, as a starting point, we recruit workers through the paid platform Amazon's Mechanical Turk (AMT); ultimately, however, we would like to transition to a more community-based method. It has been shown that, by offering an online course, one can attract a crowd of thousands or more students, who can be leveraged to improve the course itself (Weld et al., 2012). This is a skilled crowd, one which has at least the basic skill sets and prerequisites for the course. The use of such crowd workers is common in educational technology, often in a way that leverages the workers' content knowledge (Anderson, 2011; Moore et al., 2020b). Recently, crowdsourcing has become increasingly popular for content development in the education domain (Paulin and Haythornthwaite, 2016).

Companies such as Coursera and Duolingo also employ this approach to have their learners make improvements to the content as they use it. As another example, the process of crowdsourcing data from learners, or learnersourcing, has been used to identify which parts of lecture videos are confusing (Kim et al., 2014), and to describe the key instructional steps and subgoals of how-to videos (Kim, Miller, and Gajos, 2013). More recently, learnersourcing contributed to not only the annotation of existing educational content, but also the creation of novel content itself. In particular, (Williams et al., 2016) explored a crowdsourcing-based strategy towards personalized learning in which learners were asked to author paragraphs of text explaining how to solve statistics problems. The explanations generated by learners were found to be comparable in both learning benefit and rated quality to explanations produced by expert instructors.

As the fields of natural language processing and text mining continue to advance, they have seen increasingly wider usage in education technology to help automate intensive tasks (Brack et al., 2020). Previous works have looked at using different machine learning models (Pardos and Dadu, 2017; Patikorn et al., 2019) and utilizing a search engine (Karlovčec, Córdova-Sánchez, and Pardos, 2012) 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 (Slater et al., 2017). 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 (Lee et al., 2017). Much of the work in this space is focused on predicting KCs for content, after being trained on similar KC-tagged problems. Few studies have attempted to leverage text mining techniques to generate KCs for content, with no training or prediction modeling involved. A previous study made use of the same two topic modeling techniques this study utilizes, but for the context of classifying programming problems (Intisar et al., 2019). They found a series of beneficial trade-offs between the models in terms of accuracy and dimensionality, which we build upon with our work.

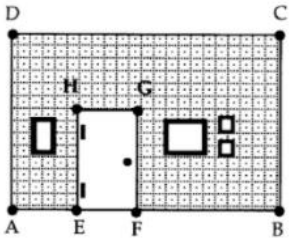
## 3. Methods

The study consists of two related experiments that differ in their domain content for a specific part of the task. The first domain is mathematics, with a focus on the area of shapes, such as squares and rectangles. The second domain is English writing, with a focus on prose style involving agents and clause topics. The math content is at the middle school level and the writing content at the undergraduate level in the United States. For both domains, we conducted an experiment using the Amazon Mechanical Turk (AMT) platform. Eighty unique crowd workers on AMT, known as turkers, completed the math experiment and sixty unique turkers completed the writing, for a total of 140 participants. Thirteen participants in the Writing experiment were removed from our analyses due to submitting invalid responses that indicated either a complete misunderstanding of the experiment instructions or behavior similar to a bot. Filtering these invalid participants left us with 47 total participants in writing, combined with the 80 from math, for a total of 127 participants. Among these participants, 55 self-identified as female and 72 as male. The mean selected age range was 35-44. All participants reported having at least a high school degree or equivalent. Additionally, none of the participants indicated that their occupation directly involved the use of geometry or writing prose. In each experiment, the tasks took roughly five minutes to complete. Participants were paid $0.75 on completion, providing a mean hourly wage of $9.

In the main task of the experiment, participants were given a word problem and asked to list three KCs that are required to solve the problem. The math experiment problem is about finding the area of a shape with two structures, while the writing experiment problem involves the revision of a sentence (Figure 1 and 2). The prompt for KCs was “As concisely as possible, please indicate a skill required to answer the above math problem about the wall.” in the math experiment and “As concisely as possible, please indicate a skill required to answer the above question that involved revising the sentence.” in the writing experiment. Note that the prompt uses the term “skill” rather than “knowledge component” to avoid jargon that may be confusing.

### *3.1 Math and Writing Experiments*

The math word problem for which participants were asked to generate KCs is shown on the left side of Figure 1. This problem, along with the two priming questions, comes from a dataset titled ‘Geometry Area (1996-97) [KRM]’, which was used in a previous study of a geometry cognitive tutor (Stamper and Koedinger, 2011). An expert instructor familiar with the domain knowledge and KC modeling process tagged this problem with three KCs as described in Table 1. These will serve as a baseline for our comparison with the turkers’ generated KCs. The word problem for which participants were asked to generate KCs is shown on the right side of Figure 1. This problem comes from an online prose style course for freshmen and sophomores at a four-year university in the United States. Similar to the math experiment, we use four KCs provided by an expert instructor (Table 1) as a baseline for comparison




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. _____

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.*

Your answer: _____

*Figure 1*: The math and writing word problems that participants provided three skills for in both conditions in either the math or writing experiment.

Table 1. *Expert-generated KCs in the math and writing experiment, domain code “M” for math and “W” for writing.*

| **KC (Domain)** | **Definition** |
|---|---|
| compose-by-addition (M) | In an equation such as a + b = c, given any two of a, b or c, find the third variable. |
| subtract (M) | Subtract the area of one shape from another. |
| rectangle-area (M) | Finding the area of a rectangle shape. |
| id-clause (W) | Identify the clause-level topic of a sentence. |
| discourse-level (W) | Keep the discourse-level topic of the sentence in focus. |
| subject-position (W) | Assess whether an entity is a subject. |
| verb-form (W) | Transform a passive verb to an active verb. |

### *3.2 Topic Modeling & Evaluation*

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 the text. To create clusters that were easier to quickly process and identify the common themes of the crowdworker contributions, we leveraged topic modeling. The crowdsourced KC text was lemmatized and stop words were removed, using a common NLP library in Python (Bird, Klein, and Loper, 2009). No further text processing was performed on the crowdsourced KC data before running them through the topic models, as we wanted results without fine-tuning any parameters or heavily processing the data. To support replicability, the code used to generate both the topic models used in this study can be found at: https://github.com/StevenJamesMoore/ICCE2020/blob/master/topic_models.ipynb.

The first topic modeling technique we used is Latent Dirichlet Analysis (LDA (Blei, Ng, and Jordan, 2003)). LDA maps all documents, in this case the participant-generated KCs for the problems, to a set number of topics in a way such that the words in each document are captured by the topics (AlSumait et al., 2009). For this process we set the number of topics to five, as the math problem has three expert-generated KCs and the writing problem has four expert-generated KCs. We wanted a topic number similar to the actual number of expert KCs, but not limited to it, in case participants contributed potentially novel or latent KCs of the problems. Additionally, in this case the documents across our corpus are small in length, often only consisting of a few words or phrases. This is not a traditional document used by LDA, which generally has a topic characterized by a distribution over words (Yau et al., 2014). However, the model is still able to label skills from the brief text and identify relationships between the words that prove useful (Tong and Zhang, 2016).

Non-negative Matrix Factorization (NMF (Lee and Seung, 2001)) was the second topic modeling technique we used, which conceptually outputs topics made up of term clusters from the documents fed into it. NMF uses linear algebra for topic modeling by identifying the latent structure in data, the participant-generated KCs for the problems, represented as a non-negative matrix (Luo et al., 2017). Again, the topic number for this model was also set at five and the other parameters of the code can be accessed via the GitHub link. We chose this for the second topic modeling method as it successfully has been used in several related educational contexts previously, in order to map skills to assessment items and improve Q-matrices (Desmarais, 2012; Desmarais and Naceur, 2013).

The results of the topic models were then evaluated by two professional instructional designers who read the corresponding topic terms and also had access to the problem that crowdworkers provided KCs for, but they were not made aware of the expert-generated KCs that were previously used for the problems. These two experts worked collaboratively to construct interpretations of the topics, based on the topic five key terms that comprised them. The researchers were then tasked with identifying if these topic interpretation labels correspond to any of the expert-generated KCs for the problems. Therefore, the topic model outputs were interpreted by the researchers and compared to the KCs generated by experts. This was the evaluation phase of comparing the topic model outputs, ran on the crowdsourced KCs, to the ground truth expert-generated KCs for the math and writing problems.

## 4. Results

The five topics from both the LDA and NMF models for the math experiment data, along with the top five most common terms associated with each topic, are presented in Table 2. This table contains a topic label column, which was constructed as an interpretation of terms by the two instructional designers during the evaluation phase. The two models share similar topics and the top five terms that comprise them are also comparable. Both models include topics that involve basic arithmetic operations used in the problem, such as addition, subtraction, and multiplication. They differ in that the LDA one contains a topic mentioning the order of operations, which is applicable to the problem. The NMF model also contains two unique topics that mention calculations with double digit numbers and another that includes division. Interestingly, division is not used in the problem, while the problem does involve order of operations.

Table 2. *Top 5 terms from 5 topics identified by the LDA and NMF models from the math data*

| Topic # | Topic Terms | Topic Label |
|---|---|---|
| **LDA Model** | | |
| 1 | wall, calculation, _num, read, equation | Perform a calculation using the wall values |
| 2 | area, door, rectangle, subtract, wall | Subtracting the area of the door from the wall |
| 3 | multiplication, order, know, operation, subtract | Knowing to perform order of operations |
| 4 | number, add addition, subtract, digit | Adding & subtracting numbers |
| 5 | multiply, subtraction, number, digit, double | Multiplying & subtracting double digit numbers |
| **NMF Model** | | |
| 1 | number, multiply, add, digit, subtract | Multiplying & subtracting double numbers |
| 2 | multiplication, skill, determine, double, digit | Multiplying double digit numbers |
| 3 | subtraction, skill, digit, length, width | Subtraction using the length and width of a shape |
| 4 | addition, skill, word, foot, division | Adding and dividing numbers |
| 5 | area, subtract, door, know, rectangle | Subtracting the door area from the wall |

The five topics from both the LDA and NMF models for the writing experiment data, along with the top five most common terms associated with each topic, are presented in Table 2. This table also contains the topic label constructed during the evaluation phase. The topic terms between the two models are more diverse than those in the math experiment, although there is still overlap between the topics, as expected. Both models share a topic indicative of identifying the position of the sentence's subject and agent. Relatedly, they also share a topic of understanding the structure of a sentence and an understanding of how to write. Interestingly, the LDA model's topic for understanding how to write also includes mention of the english language, which is not present in the NMF model's corresponding topic. Another key topic difference is that the NMF model has two unique topics of identifying the passive voice of a sentence and has critical thinking and problem solving skills. The final noticeable topic difference is that the LDA model has a topic indicative of understanding verb tenses, which is not present in the NMF one.

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

| Topic # | Topic Terms | Topic Label |
|---|---|---|
| | **LDA Model** | |
| 1 | agent, know, subject, identify, position | Identifying the subject and agent's position |
| 2 | verb, tense, knowledge, understand, know | Understanding verb tenses |
| 3 | sentence, structure, understand, construction, determine | Understanding the structural components of a sentence |
| 4 | english, skill, understand, language, writing | Understanding the English language and how to write it |
| 5 | comment, writing, english, mean, language | Understanding the English language and how to write it |
| | **NMF Model** | |
| 1 | english, comment, voice, passive, interpret | Identifying the passive voice of a sentence |
| 2 | know, agent, subject, position, identify | Identifying the subject and agent's position |
| 3 | sentence, knowledge, structure, determine, meaning | Understanding the structural components of a sentence |
| 4 | skill, writing, paragraph, require, word | Knowing how to write |
| 5 | problem, thinking, critical, solve, math | Critical thinking and problem solving skills |

After the creation of the topic interpretations during the evaluation phase, the researchers matched the topics to any corresponding expert-generated KCs for both the math and writing problems. The matching of any topics to these KCs is presented in Table 4. As depicted, each domain had expert-generated KCs that were not addressed by a topic from either model.

The math experiment's first LDA model topic resembles the *compose-by-addition* KC as it mentions the calculation of values from the wall depicted in the problem. Both models mention subtracting across multiple topics, but only the second LDA topic and the fifth NFM topic address the *subtract* KC. To indicate this KC, the topic needed to explicitly mention subtracting the area of one shape (door) from another (wall or rectangle). The final math KC, *rectangle-area*, was not present in any model's topic, although it may be argued that the previous topics for the *subtract* KC also include this one implicitly.

The first two expert-generated KCs for the writing problem are not identified by any topics in either model. Both contain domain-specific vocabulary, clause and discourse, which is not included in the top five terms for the topics. Topic three for both the LDA and NMF models indicates the structure of the sentence, which is a step towards these KCs, but not inclusive enough to match either of them. Both models had a single topic that matched the *subject-position* KC, as the topics explicitly mention the position of the subject and agent in the sentence. The final expert-generated KC, *verb-form*, is partially addressed by the second LDA topic, which mentions verb tense. However, the NFM model's first topic more explicitly addresses this KC, as it specific the passive voice of the sentence.

Table 4. *Expert-generated KCs in the math (M) and writing (W) experiment domains compared against the topics generated by the two models.*

| KC (Domain) | LDA Topic # | NMF Topic # |
|---|---|---|
| compose-by-addition (M) | 1 | - |
| subtract (M) | 2 | 5 |
| rectangle-area (M) | - | - |
| id-clause (W) | - | - |
| discourse-level (W) | - | - |
| subject-position (W) | 1 | 2 |
| verb-form (W) | 2 | 1 |

## 5. Discussion

Firstly, we wanted to better understand how we could have crowdworkers assist in the generation and mapping of KCs to problems across the math and writing domains. Using the results of the evaluation from expert-generated KC to topic, crowdworkers appear to have similar performance in terms of the number of expert KCs they were able to generate. The expert-generated KCs in the math domain had little domain vocabulary or included terms that were not mentioned in the question. Only 'rectangle' was explicitly used in the KCs, which is not included in the question text, but that is also a more common word and does not require domain knowledge to know. On the other hand, all four expert-generated KCs in the writing experiment contain vocabulary specific to the problem and its domain. It is not a surprise that the first two, which include clause and discourse, were not indicated by participants. These terms are jargon and not as common, particularly for those who may not have had an English writing course recently or speak it as a secondary language. Additionally, the math domain problem comes from a middle school course while the writing domain one comes from an undergraduate college course. It was surprising at first that participants were able to match the subject-position KC, however those terms are explicitly mentioned in the instructional text of the problem. It appears that the strategy some participants employed was using the instructions of the question to generate the KCs. This is akin to previous work that uses different machine learning models to predict KCs, which often employs the problem's text as a key feature (Karlovčec et al., 2012).

Secondly, we wanted to see how topic modeling could be leveraged with crowdsourced KCs to produce results comparable to that of experts. This process yields a large amount of text data, that is both a mixture of applicable, redundant, and irrelevant all at the same time. To help make sense of this data in a less time-consuming matter, we turned to topic modeling techniques to aggregate the data. We found that the topics these models produce are indicative of several KCs for the problems that are akin to one's experts previously generated and used for them. In the math experiment, two out of three of the expert KCs were produced between both models. In the writing experiment, another two of the expert KCs were produced out of four in total. Ultimately both models across each domain were comparable, resulting in almost identical topics that related back to an expert-generated KC. The LDA model for math had a slight gain over the NMF one, in that it had a topic comparable to another KC that the NMF one did not. However, this was perhaps the weakest topic-to-KC matching, so we cannot claim LDA performs better than NMF or vice-versa for this task.

The difference in topics between the two models for the math domain were not useful in the identification of an expert-generated KC. These different topics between the two addressed aspects that were still relevant to the problem, such as double-digit numbers and order of operations, but they were not the specified KCs for the problem. This was also true for the topics both math models shared, while they were mostly applicable to the problem due to indicating multiplication and addition, they were not relevant KCs. The LDA and NFM models in the writing domain had similar results, in that the topics that differed between the two pertained to the problem, but were not a KC for it. Interestingly, the LDA model for the writing domain was the only model that had a duplicate topic within its own top five topics, which was about understanding and writing the English language. While that may suggest participants thought such a skill was applicable to the problem, and it is, it did not match an expert-generated KC. While that is a relevant skill needed for the problem, it is not a KC the problem intends to measure, as the target learner is that in an undergraduate writing course.

### *5.1 Limitations*

In our study, we utilized two forms of topic modeling, LDA and NMF, to group potential KCs contributed by crowdworkers. Different techniques for topic modeling may produce more accurate results or ones that are more interpretable by humans (Chang et al., 2009). While performing this modeling on the data helps to aggregate it into a more manageable fashion, human interpretation is still heavily required to make use of the data. Additionally, to keep the data collection brief, participants did not receive explicit instruction on the desired granularity or process of generating KCs. Measuring a participant's existing domain knowledge and training them on this task could produce more reliable results (Koedinger and Nathan, 2004). Finally, while we had multiple researchers and experts involved in the process of interpreting the topics and evaluating their matches to expert-generated KCs, there is still a subjective element at play with what topic or KC a contribution may fall under. Making use of existing text mining techniques for this may help to increase replicability of this process (Wiedemann, 2016), along with including more data coders in this process.

## 6. Conclusion and Future Work

In this work, we solicited the knowledge components of a mathematics problem and an English writing problem from crowdworkers with different expertise and backgrounds. Our results indicate that crowdworkers can be effectively leveraged to assist in the process of generating knowledge components for problems in our math and writing domains. To our knowledge, this work is among the first to investigate the use of crowd workers to generate knowledge components for assessment activities. We found that roughly half of the KCs generated by experts and previously used for the problems in an educational setting were able to be matched by KCs generated by the crowdworkers. The LDA and NMF topic models created groupings that when interpreted by a human, were akin to the expert generated ones. Several of these topics proved useful, as they indicated KCs of the problems, but work remains to improve both interpretability and matching more domain-specific KCs. Ultimately the results of this work support the benefits of engaging the crowd in more nuanced learning science tasks, as well as encourage the adoption of a similar process that utilizes learnersourcing. With continued improvement of the efficiency and efficacy of this process, we envision further use of crowdworkers to generate high quality and accurate knowledge components that easily scale up to benefit both learners and education providers.

Building upon this work, we plan to integrate this process in a learnersourced context, where participants (i.e., students) potentially have more commitment and domain knowledge that could be leveraged (Paulin and Haythornthwaite, 2016). This would enable us to properly train them to provide such KCs 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 KCs for problems; these problems are then peer reviewed and presented to the teachers (or relevant parties) to help them confirm applicable 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 (Bernstein, 2013). This study demonstrates the first step in developing such a workflow, providing initial insights into how crowdsourced contributions might be leveraged in the KC mapping process. While the analyses we performed here may end up taking more time than the task of generating the KCs themselves, they contribute to an eventual domain-independent workflow that can greatly expedite such tasks.

Our findings also suggest several directions for future research in crowdsourcing KCs and involving crowdworkers in instructional design tasks without the need for prior training. Performing a similar study involving different domains at various grade levels might yield interesting results as well, especially when crowdworkers might have more familiarity thanks to their diverse backgrounds. Additionally, studying the impact of grouping crowdworkers on their reported expertise levels and then having them generate KCs for problems of varying difficulty levels could introduce useful design decisions for implementing this workflow into a digital learning environment. Ultimately, we would like to scale this up into a learnersourced context, such as embedding these prompts into an online course. Our next step is to leverage different natural language processing techniques outside of topic modeling, such as keyphrase extraction or summarization, in order to achieve higher accuracy and reduce the need for human input into the interpretability aspect of this work.

## References


AlSumait, Loulwah, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. "Topic Significance Ranking of LDA Generative Models." Pp. 67–82 in *Joint European Conf. on ML and KDD*. Springer.

Anderson, Michael. 2011. "Crowdsourcing Higher Education: A Design Proposal for Distributed Learning." *MERLOT Journal of Online Learning and Teaching* 7(4):576–590.

Bernstein, Michael S. 2013. "Crowd-Powered Systems." *KI-Künstliche Intelligenz* 27(1):69–73.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022.

Brack, Arthur, Jennifer D'Souza, Anett Hoppe, Sören Auer, and Ralph Ewerth. 2020. "Domain-Independent Extraction of Scientific Concepts from Research Articles." *ArXiv Preprint ArXiv:2001.03067*.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." Pp. 288–296 in *Advances in neural information processing systems*.

Cooper, Seth, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, and Zoran Popović. 2010. "Predicting Protein Structures with a Multiplayer Online Game." *Nature* 466(7307):756–760.

Corbett, Albert T., and John R. Anderson. 1994. "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge." User Modeling and User-Adapted Interaction 4(4):253–278.

Desmarais, Michel C. 2012. "Mapping Question Items to Skills with Non-Negative Matrix Factorization." *ACM SIGKDD Explorations Newsletter* 13(2):30–36.

Desmarais, Michel C., and Rhouma Naceur. 2013. "A Matrix Factorization Method for Mapping Items to Skills and for Enhancing Expert-Based q-Matrices." Pp. 441–450 in *International Conference on Artificial Intelligence in Education*. Springer.

Intisar, Chowdhury Md, Yutaka Watanobe, Manoj Poudel, and Subhash Bhalla. 2019. "Classification of Programming Problems Based on Topic Modeling." Pp. 275–283 in *Proceedings of the 2019 7th International Conference on Information and Education Technology*.

Karlovčec, Mario, Mariheida Córdova-Sánchez, and Zachary A. Pardos. 2012. "Knowledge Component Suggestion for Untagged Content in an Intelligent Tutoring System." Pp. 195–200 in *International Conference on Intelligent Tutoring Systems*. Springer.

Kim, Juho, Philip J. Guo, Daniel T. Seaton, Piotr Mitros, Krzysztof Z. Gajos, and Robert C. Miller. 2014. "Understanding In-Video Dropouts and Interaction Peaks Inonline Lecture Videos." Pp. 31–40 in *Proceedings of the first (2014) ACM conference on Learning@ scale conference*.

Kim, Juho, Robert C. Miller, and Krzysztof Z. Gajos. 2013. "Learnersourcing Subgoal Labeling to Support Learning from How-to Videos." Pp. 685–690 in *CHI'13 Extended Abstracts on Human Factors in Computing Systems*.

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

Koedinger, Kenneth R., Vincent Aleven, Neil Heffernan, Bruce McLaren, and Matthew Hockenberry. 2004. "Opening the Door to Non-Programmers: Authoring Intelligent Tutor Behavior by Demonstration." Pp. 162–174 in International Conference on Intelligent Tutoring Systems. Springer.

Koedinger, Kenneth R., Albert T. Corbett, and Charles Perfetti. 2012. "The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning." *Cognitive Science* 36(5):757–798.

Koedinger, Kenneth R., Elizabeth A. McLaughlin, and John C. Stamper. 2012. "Automated Student Model Improvement." International Educational Data Mining Society.

Koedinger, Kenneth R., and Mitchell J. Nathan. 2004. "The Real Story behind Story Problems: Effects of Representations on Quantitative Reasoning." *The Journal of the Learning Sciences* 13(2):129–164.

Koedinger, Kenneth R., John C. Stamper, Elizabeth A. McLaughlin, and Tristan Nixon. 2013. "Using Data-Driven Discovery of Better Student Models to Improve Student Learning." Pp. 421–430 in *International Conference on Artificial Intelligence in Education*. Springer.

Lee, Daniel D., and H. Sebastian Seung. 2001. "Algorithms for Non-Negative Matrix Factorization." Pp. 556–562 in *Advances in neural information processing systems*.

Lee, Tak Yeon, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. "The Human Touch: How Non-Expert Users Perceive, Interpret, and Fix Topic Models." *International Journal of Human-Computer Studies* 105:28–42.

Liu, Ran, and Kenneth R. Koedinger. 2017. "Closing the Loop: Automated Data-Driven Cognitive Model Discoveries Lead to Improved Instruction and Learning Gains." *Journal of Educational Data Mining*

9(1):25–41.
Liu, Ran, Elizabeth A. McLaughlin, and Kenneth R. Koedinger. 2014. “Interpreting Model Discovery and Testing Generalization to a New Dataset.” in *Educational Data Mining 2014*. Citeseer.
Luo, Minnan, Feiping Nie, Xiaojun Chang, Yi Yang, Alexander Hauptmann, and Qinghua Zheng. 2017. “Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling.” in *Thirty-first AAAI conference on artificial intelligence*.
Moore, Steven, and John Stamper. 2019. “Decision Support for an Adversarial Game Environment Using Automatic Hint Generation.” Pp. 82–88 in *International Conference on Intelligent Tutoring Systems*. Springer.
Moore, Steven, Nguyen, Huy, and John Stamper. 2020a. “Evaluating Crowdsourcing and Topic Modeling in Generating Knowledge Components from Explanations.” Pp. 398–410 in *International Conference on Artificial Intelligence in Education*. Springer.
Moore, Steven, Nguyen, Huy, and John Stamper. 2020b. “Towards Crowdsourcing the Identification of Knowledge Components.” Pp. 245–248 in *Proceedings of the Seventh (2020) ACM Conference on Learning@ Scale*.
Nguyen, Huy, Yeyu Wang, John Stamper, and Bruce M. McLaren. 2019. “Using Knowledge Component Modeling to Increase Domain Understanding in a Digital Learning Game.” in *International Educational Data Mining Society*.
Pardos, Zachary A., and Anant Dadu. 2017. “Imputing KCs with Representations of Problem Content and Context.” Pp. 148–155 in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*.
Patikorn, Thanaporn, David Deisadze, Leo Grande, Ziyang Yu, and Neil Heffernan. 2019. “Generalizability of Methods for Imputing Mathematical Skills Needed to Solve Problems from Texts.” Pp. 396–405 in *International Conference on Artificial Intelligence in Education*. Springer.
Paulin, Drew, and Caroline Haythornthwaite. 2016. “Crowdsourcing the Curriculum: Redefining e-Learning Practices through Peer-Generated Approaches.” *The Information Society* 32(2):130–142.
Schraagen, Jan Maarten, Susan F. Chipman, and Valerie L. Shalin. 2000. *Cognitive Task Analysis*. Psychology Press.
Slater, Stefan, Ryan Baker, Ma Victoria Almeda, Alex Bowers, and Neil Heffernan. 2017. “Using Correlational Topic Modeling for Automated Topic Identification in Intelligent Tutoring Systems.” Pp. 393–397 in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*.
Stamper, John C., and Kenneth R. Koedinger. 2011. “Human-Machine Student Model Discovery and Improvement Using DataShop.” Pp. 353–360 in *International Conference on Artificial Intelligence in Education*. Springer.
Stamper, John, Kenneth Koedinger, and Elizabeth Mclaughlin. 2013. “A Comparison of Model Selection Metrics in Datashop.” in Educational Data Mining 2013.
Tong, Zhou, and Haiyi Zhang. 2016. “A Text Mining Research Based on LDA Topic Modelling.” Pp. 21–22 in *Proceedings of the Sixth International Conference on Computer Science, Engineering and Information Technology (CCSEIT)*.
Weld, Daniel S., Eytan Adar, Lydia Chilton, Raphael Hoffmann, Eric Horvitz, Mitchell Koch, James Landay, Christopher H. Lin, and Mausam Mausam. 2012. “Personalized Online Education—a Crowdsourcing Challenge.” in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.
Wiedemann, Gregor. 2016. “Integrating Text Mining Applications for Complex Analysis.” Pp. 55–166 in *Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*, *Kritische Studien zur Demokratie*, edited by G. Wiedemann. Wiesbaden: Springer Fachmedien.
Williams, Joseph Jay, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. “Axis: Generating Explanations at Scale with Learnersourcing and Machine Learning.” Pp. 379–388 in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*.
Yau, Chyi-Kwei, Alan Porter, Nils Newman, and Arho Suominen. 2014. “Clustering Scientific Documents with Topic Modeling.” *Scientometrics* 100(3):767–786.
Yuen, Man-Ching, Irwin King, and Kwong-Sak Leung. 2011. “A Survey of Crowdsourcing Systems.” Pp. 766–773 in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE.