Toward Personalizing Students' Education with Crowdsourced Tutoring

Ethan Prihar
Worcester Polytechnic Institute
ebprihar@wpi.edu

Thanaporn Patikorn
Worcester Polytechnic Institute
tpatikorn@wpi.edu

Anthony Botelho
Worcester Polytechnic Institute
abotelho@wpi.edu

Adam Sales
Worcester Polytechnic Institute
asales@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

ABSTRACT

As more educators integrate their curricula with online learning, it is easier to crowdsource content from them. Crowdsourced tutoring has been proven to reliably increase students' next problem correctness. In this work, we confirmed the findings of a previous study in this area, with stronger confidence margins than previously, and revealed that only a portion of crowdsourced content creators had a reliable benefit to students. Furthermore, this work provides a method to rank content creators relative to each other, which was used to determine which content creators were most effective overall, and which content creators were most effective for specific groups of students. When exploring data from TeacherASSIST, a feature within the ASSISTments learning platform that crowdsources tutoring from teachers, we found that while overall this program provides a benefit to students, some teachers created more effective content than others. Despite this finding, we did not find evidence that the effectiveness of content reliably varied by student knowledge-level, suggesting that the content is unlikely suitable for personalizing instruction based on student knowledge alone. These findings are promising for the future of crowdsourced tutoring as they help provide a foundation for assessing the quality of crowdsourced content and investigating content for opportunities to personalize students' education.

Author Keywords

Online Tutoring; Crowd Sourcing; Statistical Analysis; Personalized Education

L@S'21, June 22–25, 2021, Virtual Event, Germany.

ACM ISBN 978-1-4503-8215-1/21/06.
https://doi.org/10.1145/3430895.3460130

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); Empirical studies in HCI;

INTRODUCTION

The need for crowdsourcing within online learning platforms is growing as the user base of these platforms continues to expand and diversify [18, 7]. Crowdsourcing can be used effectively to generate new teaching materials [22] and new tutoring for students [18]. As more platforms integrate crowdsourcing, methods to evaluate and maintain the quality of crowdsourced materials need to be developed to ensure students receive a high quality education and effective support.

In the 2017-2018 academic year, ASSISTments, an online learning platform [10], deployed TeacherASSIST. TeacherASSIST allowed teachers to create tutoring in the form of hints and explanations for problems they assigned to their students. TeacherASSIST then redistributed teachers' tutoring to students outside of the their class. At L@S 2020, ASSISTments reported that teachers created about 40,000 new instances of tutoring for about 26,000 different problems. Through two large-scale randomized controlled experiments, it was determined that there was statistically significant improvement on the next problem correctness of students who received crowdsourced tutoring. Since the publication of these findings, ASSISTments has scaled up the distribution of crowdsourced content within the platform. The first part of this study uses new data, collected from the 2019-2020 and 2020-2021 school years to re-evaluate the findings of the original study and confirm that crowdsourced tutoring continues to benefit students overall.

The second part of this study investigated if there was a significant difference between the quality of different teachers' tutoring. The methodology used in this paper could be used in the future to determine which teacher's content should have priority when distributing tutoring to students in other classes.

Lastly, this study determined if there were any qualitative interactions between the teachers who created tutoring and

students grouped by their knowledge-level. Personalized learning requires qualitative interactions, defined as one group of students benefiting more from one type of instruction, while a different group of students benefited more from an alternative type of instruction. The learning science community has spent a considerable amount of time investigating the impact of personalized learning on students. While personalized tutoring based on prior knowledge has shown some evidence of a qualitative interaction [20], other methods for personalization, such as learning styles, have rarely shown conclusive evidence of a qualitative interaction [17]. The method used in this study can be used to search experimental data for qualitative interactions without using a randomized controlled trial to directly evaluate the presence of a particular qualitative interaction.

Specifically, this work seeks to address the following research questions:

1. Do the findings of the previous TeacherASSIST study still hold when tested on new data?
2. How did the effectiveness of teachers' tutoring compare to each other?
3. Was there any potential to personalize the tutoring students received based on their knowledge-level?

BACKGROUND

The Value of Crowdsourcing

The growing popularity of online learning platforms has created a greater opportunity and a greater need for educational materials of all levels. With a greater diversity of students, there arises the need to provide instruction to students of varying skill levels. Crowdsourcing can help diversify the available tutoring and assist in personalizing lesson plans for students [25, 3]. Crowdsourcing offers a mechanism to obtain the breadth of educational content required to meet the growing demand of online tutoring, but poses some challenges as well [25]. The biggest risk from using crowdsourced materials is the potential for low quality, or misleading material to negatively impact students [25]. Even if the information is high-quality, overly detailed tutoring, or tutoring from highly different sources can also have a negative impact of students' learning [23, 13, 12]. Ways to mitigate these risks include algorithmically evaluating the quality of crowdsourced content creators [21], or simply crowdsourcing content only from people that have been deemed qualified [16, 4, 24, 5].

Even with these risks, crowdsourcing has been a viable method for obtaining information on the knowledge components of different math problems [15], assisting students learning computer programming [2], and collecting videos explaining how to solve mathematics problems [26, 27]. Most directly, in the study preceding this work, tutoring messages created by teachers, for students completing work in ASSISTments, had an overall positive effect on students' learning [18]. Although crowdsourcing has shown promising results in many situations, there is a need to continue to evaluate the methods through which crowdsourced content is collected and validated so that as more educational platforms begin to incorporate crowdsourcing, they can do so efficiently, effectively, and without risk to students.

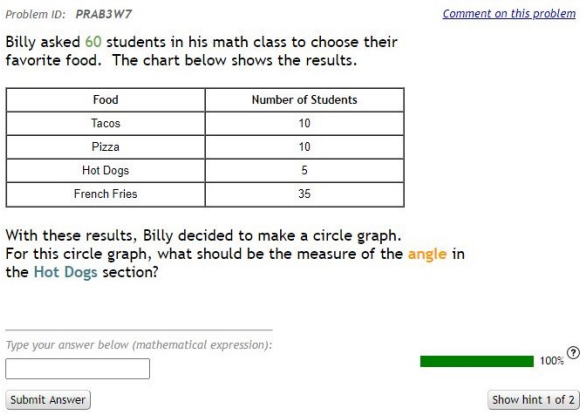

Figure 1. The ASSISTments tutor, as scene by a student solving a mathematics problem.

ASSISTments

The data used in this study comes from ASSISTments. ASSISTments[1] [11] is an online learning platform focused on empowering teachers via automating laborious tasks such as grading and record keeping of students, and providing insight to teachers on their class's common wrong answers and miss conceptions on assignments [11]. ASSISTments provides K-12 mathematics problems and assignments from multiple open source curricula for teachers to choose from and assign to their students. After an assignment has been assigned to students, students complete the assignment in the ASSISTments tutor, shown in Figure 1 [18]. In the tutor, students receive immediate feedback when they submit a response to a problem, which informs them if they are correct [9]. For some problems, students can request tutoring, which is available to them at any point during their completion of the problem, regardless of whether or not they have already attempted the problem. Tutoring comes in the form of hints, explaining how to solve parts of the problem, [11, 20], examples of how to solve similar problems [8, 14], examples of incorrect responses to problems with explanations of the error [14, 1], and full solutions to problems [27, 26]. Two examples of tutoring in ASSISTments are shown in 2 [18].

Recently ASSISTments began a program called TeacherASSIST, in which tutoring was crowdsourced from teachers in the form of written and video-recorded hints and explanations for solving middle-school math problems. ASSISTments collected tutoring created by teachers who had already used the platform for their own classrooms, and then provided the crowdsourced hints and explanations to students. Distributing these hints and explanations lead to a positive impact on students' learning [18]. In this study, the data released from the TeacherASSIST study [19], new data from TeacherASSIST collected since the publication of the previous study, and information on students' knowledge-level collected from the ASSISTments platform were used to investigate if any content creators' tutoring significantly out-performed other content creator's tutoring, as well as determine if there were any qualitative interactions between content creators and students.

[1] https://www.ASSISTments.org/

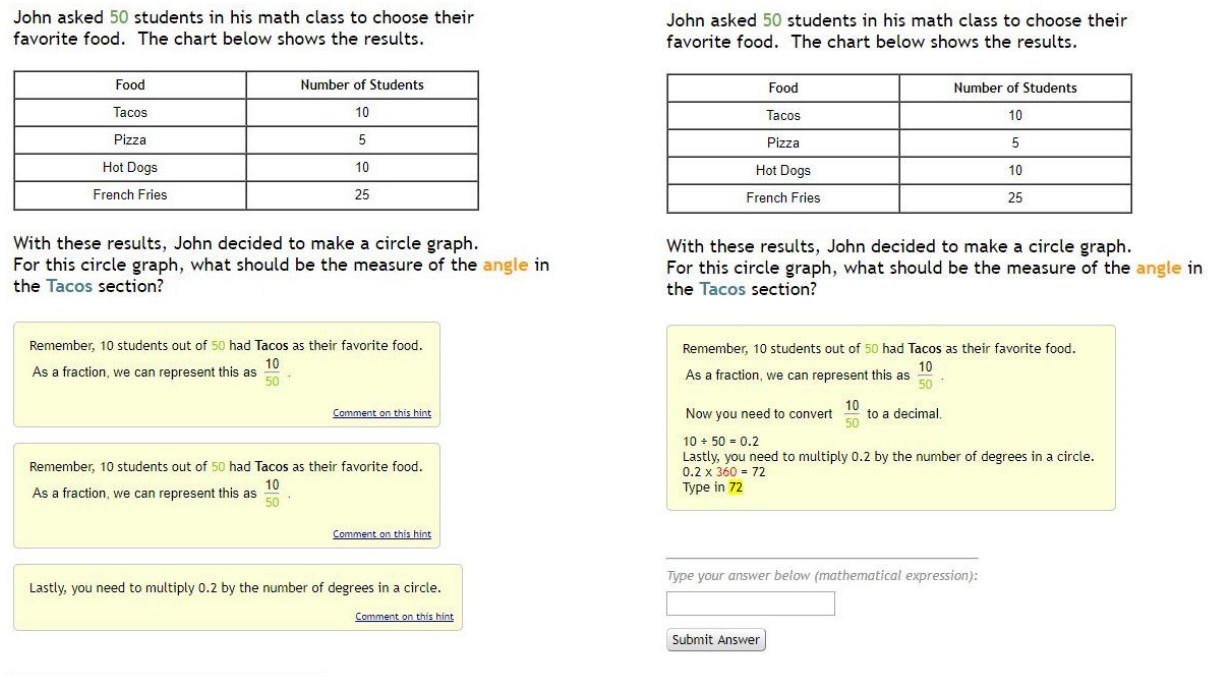

Figure 2. Two instances of tutoring in ASSISTments. On the left is a series of hints. On the right is a full explanation of how to solve the problem.

METHODOLOGY

Confirming the Previous Study's Findings

The same analysis performed in the original study [18] was repeated using the exact same code from the previous study made available by the Open Science Foundation [19]. New data, collected since the completion of the previous study up until February 2, 2021, was used to determine if the previously reported positive impact of TeacherASSIST was still present in a new academic year. The new dataset contained 6,774 unique problems, 7,059 unique tutoring messages, 18,420 unique students, and 500,900 answered problems. 50,426 of the answered problems were answered by students in the control condition, where they were not given the option to request tutoring, and 450,474 of the problems were answered by students in the intent-to-treat condition, in which they had the option to, but did not necessarily request tutoring. A majority of students were placed in the treatment condition because the previous study found the treatment condition to have a reliable positive effect, and ASSISTments did not want to prevent half the students from receiving beneficial crowdsourced tutoring. Of all the students in the new dataset, only 7.92% of them appeared in the initial study's data as well.

In order to gain more insight into how reliable the findings of the initial study were, a problem-level and student-level intent-to-treat analysis, in which the students were considered to be in the treatment condition if they were given the option to receive crowdsourced tutoring, regardless of whether or not they received it, and a treated analysis, where a student was considered to be in the treatment condition only if they received crowdsourced tutoring, were performed. For all of these analyses, which were all performed in the initial study, the Benjamini-Hochberg procedure was used to control the false discovery rate [6].

Measuring the Effectiveness of Teachers

To determine the effectiveness of each teacher, the data from the previous study and this study were combined and filtered such that only the instances where a student received no tutoring, or crowdsourced tutoring **for the first time**, and then immediately answered another problem remained. This step was necessary to remove compounding and extended exposure effects that would occur if students' next problem correctness was used to evaluate the quality of teacher's tutoring after students had seen tutoring from multiple teachers. Furthermore, any teachers whose tutoring was only seen by fewer than 30 students was excluded, as there was insufficient data to measure the effectiveness of these teachers. After data processing, 31,616 instances of a student getting one of 1,026 problems wrong, receiving tutoring from one of 11 different teachers, and then answering one of 1,308 different problems were used in the following analysis.

The filtered data was used to fit a regression which predicted next problem correctness based on the student, the problem the student got wrong, the teacher who wrote the tutoring that the student saw upon getting the problem wrong, and the next problem used to evaluate the quality of the tutoring. In addition to accounting for compounding and extended exposure effects, the students, and the problems they completed, were abstracted into sets of representative features. The features for students are shown in Table 1, and the features for problems are shown in Table 2. These features were used in the model instead of unique identifiers for each student and problem for two reasons. Primarily, using features to represent students and

Student Features
Total number of problems answered
Mean correctness on all completed problems
Mean time until first response on all completed problems
Mean time on task per problem
Mean number of attempts per problem

Table 1. Features used to abstract students while measuring the effectiveness of teacher's tutoring.

Problem Features
Type of problem, e.g., multiple choice, algebraic response
Mean correctness of all answers submitted for the problem
Mean time until first response for all students that answered the problem
Mean time on task of all answers submitted for the problem
Mean number of attempts of all answers submitted for the problem

Table 2. Features used to abstract problems while measuring the effectiveness of teacher's tutoring.

problems makes it easier to generalize this procedure to other data from different educational platforms. Secondly, given the large number of unique students and problems, a model trained to predict next problem correctness would likely over-fit and obtain very high accuracy by recognizing unique combinations of students and problems, rather than estimating correctness based on the teacher who created the tutoring given to the student, as intended.

Unlike the students and problems, teachers were not abstracted into representative features, as the goal of this process was to evaluate the effectiveness of the individual teachers, not the effectiveness of the different qualities of teachers. Teacher's unique identifiers were one-hot encoded for use in the model. In cases from the control condition, where students did not receive tutoring, all of the one-hot encoded teacher covariates equaled zero. By structuring the model's inputs this way, the coefficient of each teacher covariate measured how much more or less likely a student was to get the next problem correct after receiving tutoring from the corresponding teacher, and the probability of the null hypothesis for the covariate was the probability that receiving tutoring from the corresponding teacher was not better than receiving no tutoring at all. The probability of the null hypothesis was adjusted using the Benjamini-Hochberg procedure for controlling the false discovery rate [6] because each determination of the effectiveness of a teacher's tutoring was treated as a separate hypothesis. This model was used to determine which teachers' tutoring was statistically significantly better for students than receiving no tutoring.

Comparing the Effectiveness of Different Teachers

In addition to using the model from the previous section to evaluate the overall effectiveness of each teacher's tutoring, the model can also be used to compare teachers to each other. Comparing the coefficient of each teacher to determine which teacher's tutoring has a larger treatment effect is, alone, not enough to confirm that one teacher's tutoring is truly more effective that another teacher's tutoring, as the standard deviation of the difference between the teachers' effectiveness could be so large that the difference between the teachers' coefficients is statistically insignificant. However, using the variance-covariance matrix, the standard deviation of the difference between two teachers' coefficients can be calculated using Equation 1, where $var(T_x)$ is the variance of teacher x's coefficient, $var(T_y)$ is the variance of teacher y's coefficient, $cov(T_x, T_y)$ is the covariance of teacher x's and y's coefficient from the variance-covariance matrix, and δ is the standard deviation of the difference between teacher x's and y's coefficients. Then, if the difference in coefficients falls outside the 95% confidence interval, calculated using δ, it can be concluded that the teacher with a higher model coefficient created more effective tutoring than the teacher with a lower coefficient. This technique was used to create a map of teacher effectiveness, which could be used in the future to determine which teacher's tutoring should be given to struggling students.

$$\delta = \sqrt{var(T_x) + var(T_y) - cov(T_x, T_y)} \tag{1}$$

Measuring the Potential for Personalized Tutoring

The method described previously for comparing the effectiveness of different teacher's tutoring was also used to explore the data for opportunities for personalized tutoring. Personalizing the tutoring different groups of students receive based on the teacher that created the tutoring would only be justifiable, in this context, if three criteria are met:

1. One teacher's tutoring is more effective than another teacher's tutoring for one group of students. This can be determined using the method described in Section 3.3, using a model trained on only data from the students in the group.

2. The other teacher's tutoring is more effective for a separate group of students. This can also be determined using the method described in Section 3.3, using a model trained on only data from the other group of students.

3. Each teachers' tutoring is more effective than the control condition of receiving no tutoring for students in the group that benefits the most from the corresponding teacher. This can be determined using the method described in Section 3.2 on the data from only students in one group.

These criteria qualify the core assumption of personalized education, which is that in order for all students to attain the highest level of achievement they are capable of, different groups of students need to be provided with different content. If the above criteria are met, then in the future, personalizing student's educational content based on which teacher created the content would be justified. Otherwise, it would be more beneficial to give all students educational content from the teacher whose content led to the highest improvement in next problem correctness compared to the control condition. This work explored personalizing which teacher's tutoring a student received based on the knowledge-level of the student, determined by the students' average correctness.

Dependent Measure	Control Mean	Experiment Mean	t-Stat	p-Value
Correct First Try	0.65	0.66	-1.66	0.10
Requested Tutoring	0.20	0.19	2.61	**0.01**
Stop Out	0.01	0.01	1.08	0.28
Attempt Count	1.54	1.54	-0.74	0.46

Table 3. Problem-level paired t-test intention-to-treat analysis on student next-problem dependent variables. The number of unique problems = 5079.

Dependent Measure	Control Mean	Experiment Mean	t-Stat	p-Value
Correct First Try	0.63	0.64	-2.43	**0.02**
Requested Tutoring	0.20	0.20	3.22	**< 0.01**
Stop Out	0.01	0.01	-0.26	0.79
Attempt Count	1.59	1.59	0.52	0.60

Table 4. Student-level paired t-test intention-to-treat analysis on student next problem dependent variables. The number of unique students = 10340.

Dependent Measure	Control Mean	Experiment Mean	t-Stat	p-Value
Correct First Try	0.33	0.35	-3.09	**<0.01**
Requested Tutoring	0.55	0.51	5.10	**< 0.01**
Stop Out	0.02	0.02	-0.49	0.62
Attempt Count	1.85	1.86	-0.23	0.82

Table 5. Problem-level paired t-test treated analysis on student next problem dependent variables. The number of unique problems = 2524.

Dependent Measure	Control Mean	Experiment Mean	t-Stat	p-Value
Correct First Try	0.36	0.40	-4.27	**<0.01**
Requested Tutoring	0.51	0.46	5.70	**< 0.01**
Stop Out	0.02	0.02	-0.94	0.35
Attempt Count	1.93	1.86	2.54	**0.01**

Table 6. Student-level paired t-test treated analysis on student next problem dependent variables. The number of unique students = 3547.

RESULTS

The Effectiveness of Crowdsourcing

The results of this replication of the previous study showed the same positive findings as the previous study, but with better confidence. Specifically, students who received TeacherASSIST tutoring were more likely to be able to solve the next problem correctly on their first try than students in the control condition. When students who received tutoring did not succeed on their first attempt, they were not more likely to give up or submit many more wrong answers, and they were more likely to be able to eventually solve the problem without requesting more tutoring. With this new, larger dataset, the effect on the treated is large enough to be detected with significance in the intention-to-treat analysis. Tables 3 and 4 show the results of the problem-level and student-level intention-to-treat analysis respectively, and tables 5 and 6 show the results of the problem-level and student-level treated analysis respectively. Correct first try measures the difference in next problem correctness, requested tutoring measures the difference in how much tutoring students' requested on the next problem after receiving tutoring from TeacherASSIST, Stop Out measures the difference in students' completion of the next problem, and Attempt Count measures the difference in how many attempts students' took to answer the next problem following the tutoring they received from TeacherASSIST. The bold p-values are the significant values after correcting for multiple hypothesis testing with the Benjamini-Hochberg procedure [6]. These findings confirm the previous study's conclusion that TeacherASSIST has an overall positive effect on students' learning.

Measuring the Effectiveness of Teachers

Using the method described in Section 3.2. The next problem correctness of students after receiving a teacher's tutoring was compared to receiving no tutoring. A coefficient measuring the impact of each teacher's tutoring on students' next problem correctness, a p-value denoting the probability that this coefficient is statistically equivalent to a null treatment effect, and the total number of students who viewed the tutoring from each teacher were calculated and are shown in Table 7. If a teacher's row is bold, this indicates that their tutoring had a statistically significant impact on next problem correctness after adjusting for multiple hypothesis testing.

Interestingly, even though receiving crowdsourced tutoring had an overall positive effect on students' next problem correctness, only four of the 11 teachers' tutoring had a statistically significant positive effect. Additionally, one teacher's tutoring had a statistically significantly negative impact on student's next problem correctness. This demonstrates a potential benefit to evaluating the quality of each content creator's tutoring as it is not necessarily the case that when crowdsourced content is overall beneficial, each content creator by themselves is providing a benefit. In the future of TeacherASSIST, and in other crowdsourcing endeavors, only distributing content from teachers whose tutoring has a reliable positive effect, and tutoring from teachers whose tutoring is still of ambiguous benefit, would likely lead to higher next problem correctness for students.

Teacher ID	View Count	Coefficient	p-Value
No Tutoring	2,289		
A	95	0.0629	0.112
B	222	**-0.0724**	**0.044**
C	11,202	0.0147	0.118
D	5,340	**0.0301**	**0.005**
E	76	0.0573	0.189
F	3,671	**0.0449**	**< 0.001**
G	5,763	**0.0271**	**0.008**
H	911	**0.0396**	**0.007**
I	1,452	-0.0184	0.197
J	544	0.0046	0.819
K	51	-0.0061	0.914

Table 7. The impact, statistical significance, and view count of each teacher's tutoring on students' next problem correctness.

This evaluation of teachers' effectiveness could also be used as professional development for the teachers themselves. If a teacher's tutoring is not leading to a statistically significant increase in students' next problem correctness, the crowdsourcing platform could alert these teachers that their tutoring could use improvement and provide them with examples of other teacher's tutoring that had been shown to be effective. Then, after the teacher updates their tutoring, the platform could re-evaluate their effectiveness and report back to the teacher. This interaction with teachers could also encourage teachers that are creating highly effective tutoring to create more tutoring by reporting how many students have received their tutoring, and to what extent their tutoring has helped students beyond their classroom.

Comparing the Effectiveness of Different Teachers

Using the method described in Section 3.3, the effectiveness of each teacher's tutoring was compared to every other teacher's tutoring. Figure 3 Shows the instances, in green, when the tutoring from the teacher labeled on the row, was more effective than the tutoring from the teacher labeled on the column. A grey cell indicates that the row teacher did not create more effective tutoring than the column teacher. For clarity, the teachers were sorted by how many other teachers their tutoring was more effective than. If all the teachers could be put in order from most to least effective tutoring, then Figure 3 would have entirely green cells above the diagonal. However, this is clearly not the case. Due to the variance in the effectiveness of teachers' tutoring, no teacher's tutoring is significantly better or worse than every other teachers' tutoring.

Figure 3 shows some clear examples of teachers whose tutoring is more effective than some of the other teachers' tutoring, for example, teacher F, and teachers whose tutoring is less effective than most other teachers' tutoring, for example, teacher B. Figure 3 also shows examples of teacher's whose variance in the effectiveness of their tutoring is very high, for example, teacher K. This high variance results in no teacher significantly outperforming teacher K's tutoring, and teacher K's tutoring not significantly outperforming any other teacher's tutoring. Teacher K demonstrates the need to take into account the variance of the difference between teachers' effectiveness. One cannot assert that one teacher's tutoring is more effective than another teacher's tutoring using the model coefficients alone.

Comparing teacher's tutoring can be used to choose between potential tutoring for students when more than one option is available, but care must be taken, if implementing this at scale, to not ignore tutoring from content creators with high variance in the effectiveness of their tutoring. It could be that these content creators are new to the platform, and have either created only a few instances of tutoring, or their tutoring has not had a lot of exposure yet. Content creators with high variance should be given the benefit of the doubt, and only when a teacher's tutoring is statistically significantly better than another teacher's tutoring should the more effective tutoring be chosen for the student. When using this model to select which tutoring to give the student, the student's next problem correctness should not be included in any statistical analysis that relies on random sampling.

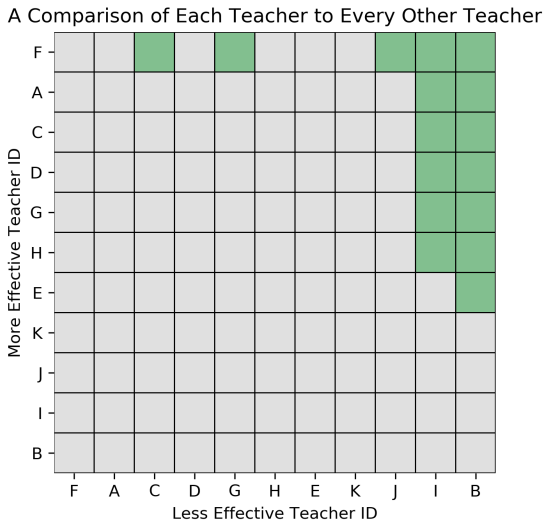

Figure 3. A map comparing the effectiveness of different teachers' tutoring.

Teacher's could also benefit from a platform that compares their effectiveness to other teachers. For professional development, teachers could be paired with a mentor and mentee. The mentor would be a teacher with statistically better tutoring than them, and the mentee would be a teacher with statistically worse tutoring than them. This would give teachers the opportunity to learn and teach others, and garner community support for the platform. Top performers could be rewarded with notoriety within the platform, and encouraged to continue to make content. Considering how heavily crowdsourcing relies on user engagement, working the analysis of teachers' effectiveness into different methods of engaging existing users and drawing in new users is an important step in the crowdsourcing process.

Measuring the Potential for Personalized Tutoring

Lastly, using the method described in Section 3.4, it was investigated if personalizing which teacher's tutoring students received based on students' knowledge-levels would likely have had a positive impact on students' next problem correctness. To group students by knowledge-level, the data was split into two datasets, The high-knowledge student data contained 18,139 instances of students whose average correctness was above average and the low-knowledge student data contained 13,475 instances of students whose average correctness was below average. To determine which teachers met Criteria 1 and 2 from Section 3.4: one teacher's tutoring is more effective than another teacher's tutoring for one group of students, and, the other teacher's tutoring is more effective for a separate group of students, the same method used in Section 3.3 was used on each group of students. The results of these comparisons are shown in Figure 4. Figure 4 shows that there is no evidence to support the claim that personalizing the tutoring students received would have led to an increase in next problem correctness. While some teachers, like teacher E, were very effective for low-knowledge students, and some teacher,

like teacher B, were particularly ineffective for high knowledge students, there were no teachers that met Criteria 2 and 3, in other words, the same teacher's tutoring was likely to have the highest positive impact on all students' next problem correctness regardless of the student's knowledge-level.

This rigorous process used to determine if there is truly a benefit to personalized tutoring could be used for more than just determining if student's tutoring can be personalized based on their knowledge-level and who created the tutoring. This process could be used on a per-problem basis. For each problem, an analysis could be performed to evaluate which of the available crowdsourced tutoring messages would be most likely to positively impact students' next problem correctness based on traits of the students. Doing this analysis on a per-problem basis would require much more data, but as platforms expand and curricula increase their integration with online learning, this may become a viable option. Additionally, if socioeconomic and demographic information on students is available, then this process could be used to personalize tutoring for students based on their gender or race. It is particularly important to pay attention to how personalization effects minority students. If the effectiveness of whatever intervention being deployed is being measured by how it effects all students on average, then in the same way that this study found that crowdsourced tutoring was overall beneficial, but some teacher's tutoring had a negative impact on next problem correctness, an intervention may be beneficial overall, but also be detrimental to minority students. Being aware of how each group of students is effected by an intervention will allow researchers to maintain fair interventions that help all students achieve their full potential.

LIMITATIONS AND FUTURE WORK

Although the results of this study are promising, there are limitations to this work. In order to compare teachers' tutoring, students and problems had to be represented with features. While these features adequately modeled students and problems well enough to account for the variations in problem difficulty and student performance, these features are not necessarily the best features to use. The features used in our models could only predict next problem correctness with an ROC AUC of 0.71. It is unlikely that the features we had available captured 100% of the variance in problems and students, and therefore including more, or different features for problems and students could increase the reliability in the measurements of the effectiveness of teacher's tutoring by increasing the model's accuracy.

In addition to potential improvements to the student and problem features, features for teachers could also be used to group teachers similar to how students were grouped in Section 4.4. Features of teachers could be used to investigate if certain groups of teachers tend to outperform other groups and could be used for personalization similarly to how individual teachers were compared in this work. Additionally, if certain features were indicative of a teacher's ability to create particularly effective tutoring, this information could be used to advise teachers and other content creators.

In this work, statistical analysis was used to determine which teachers' tutoring was most effective. While this method could be used to select which tutoring to provide to students based on which teacher is overall most effective, an online learning platform could also use reinforcement learning to select which of multiple instances of tutoring to provide to a student based on the same features of problems and students used in this work. Contextual bandit algorithms [28] use context, which in this case are features of students and problems, to take one of multiple actions, which in this case are the actions of providing one of many different instances of tutoring to a student. Then they receive a reward, which in this case would be the student's next problem correctness, and adjust their decision making process to take the action that is most likely to lead to the highest reward. While using a contextual bandit algorithm prevents one from doing the same kind of experimental analysis performed in this work, it provides a method to algorithmically determine and offer the best tutoring available to students.

Although no conclusive evidence of qualitative interactions between teachers' tutoring and students knowledge were found in this work, the potential for personalized learning should continue to be explored. More specific or alternative student features could be created evaluated for qualitative interactions the same way that knowledge-level was used in this work. It is possible that even within the dataset used in this work, there are qualitative interactions between groups of students that were not able to be considered. For example, this work had no knowledge of students' state test scores, home environments, demographic information, or socioeconomic status. All of these factors could influence what tutoring is most effective for each student and reveal the opportunity to personalize students' education.

CONCLUSION

In this follow up study, providing tutoring through TeacherASSIST continued to reliably increase students' next problem correctness, an indication that crowdsourced tutoring within the ASSISTments platform has a positive impact on students' learning. Due to many schools' recent transition to partially or fully remote learning, more data was available this year than in previous years, which allowed this study to find a reliably positive effect on students' learning even in an intent-to-treat analysis, where not every student chose to view the tutoring available to them. Furthermore, when investigating the impact of each teacher's tutoring separately, only four of the 11 teachers had a reliably positive impact on students, and one teachers' tutoring had a reliably negative impact. This finding could be used in the future to select which teacher's tutoring to provide to students based on how reliable a teachers' tutoring has been in the past. As online tutoring platforms grow and continue to incorporate crowdsourcing techniques, it will be important to include metrics for evaluating the quality of crowdsourced materials and the means to algorithmically select the most effective content. As the corpus of crowdsourced tutoring grows, the most effective content can also be explored for similarities to each other. Empirically evaluating what makes tutoring effective has the potential to improve current methods for creating tutoring, and enhance existing pedagogy.

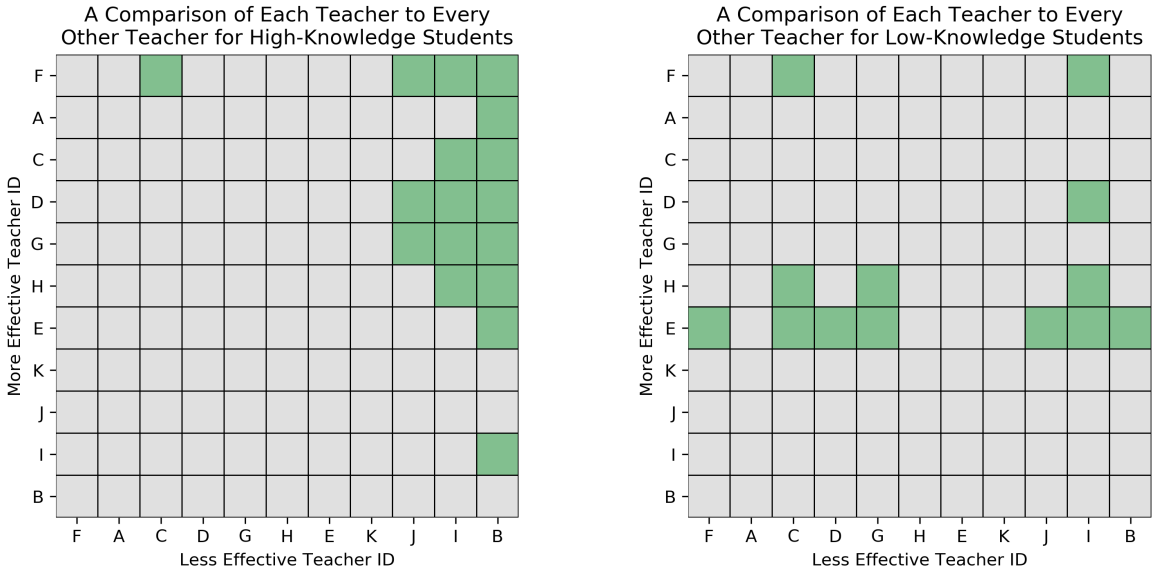

Figure 4. A map comparing the effectiveness of different teachers' tutoring separately for high and low knowledge students.

Although no evidence of the benefit of personalized education was found in this study, there is still the potential for other qualities of tutoring and the students that receive the tutoring to have an impact on what kind of tutoring is most effective. Future work can explore for more opportunities to personalize students' education using the same method in this study, or look to contextual bandit algorithms to find opportunities for personalization. Through continued efforts, crowdsourcing has the potential to advance pedagogy and provide students with a more equitable education.

ACKNOWLEDGMENTS

We would like to thank multiple NSF grants (e.g., 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), as well as the US Department of Education for three different funding lines; a) the Institute for Education Sciences (e.g., IES R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024), b) the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), and c) the EIR. We also thank the Office of Naval Research (N00014-18-1-2768), Schmidt Futures, and an anonymous philanthropic foundation.

REFERENCES

[1] Deanne M Adams, Bruce M McLaren, Kelley Durkin, Richard E Mayer, Bethany Rittle-Johnson, Seiji Isotani, and Martin Van Velsen. 2014. Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior* 36 (2014), 401–411.

[2] Dhiya Al-Jumeily, Abir Hussain, Mohammed Alghamdi, Chelsea Dobbins, and Jan Lunn. 2015. Educational crowdsourcing to support the learning of computer programming. *Research and practice in technology enhanced learning* 10, 1 (2015), 1–15.

[3] Carlos Eduardo Barbosa, Vanessa Janni Epelbaum, Marcio Antelio, Jonice Oliveira, and Jano Moreira de Souza. 2013. Crowdsourcing environments in E-learning scenario: A classification based on educational and collaboration criteria. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 687–692.

[4] Tiffany Barnes and John Stamper. 2008. Toward automatic hint generation for logic proof tutoring using historical student data. In *International Conference on Intelligent Tutoring Systems*. Springer, 373–382.

[5] Tiffany Barnes and John Stamper. 2010. Automatic hint generation for logic proof tutoring using historical data. *Journal of Educational Technology & Society* 13, 1 (2010), 3.

[6] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[7] Anthony F Botelho and Neil T Heffernan. 2019. Crowdsourcing Feedback to Support Teachers and Students. *Design Recommendations for Intelligent Tutoring Systems* 7 (2019), 101–108.

[8] Tessa HS Eysink, Ton de Jong, Kirsten Berthold, Bas Kolloffel, Maria Opfermann, and Pieter Wouters. 2009. Learner performance in multimedia learning arrangements: An analysis across instructional approaches. (2009).

[9] Mingyu Feng and Neil T Heffernan. 2006. Informing teachers live about student learning: Reporting in the assistment system. *Technology Instruction Cognition and Learning* 3, 1/2 (2006), 63.

[10] Neil T Heffernan and Cristina Lindquist Heffernan. 2014a. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 470–497.

[11] Neil T Heffernan and Cristina Lindquist Heffernan. 2014b. The ASSISTments Ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 470–497.

[12] Raymond W Kulhavy, Mary T White, Bruce W Topp, Ann L Chan, and James Adams. 1985. Feedback complexity and corrective efficiency. *Contemporary educational psychology* 10, 3 (1985), 285–291.

[13] Mary Ellen Lepionka. 2008. *Writing and developing your college textbook: a comprehensive guide to textbook authorship and higher education publishing*. Atlantic Path Publishing.

[14] Bruce M McLaren, Tamara van Gog, Craig Ganoe, Michael Karabinos, and David Yaron. 2016. The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computers in Human Behavior* 55 (2016), 87–99.

[15] Steven Moore, Huy A Nguyen, and John Stamper. 2020. Evaluating Crowdsourcing and Topic Modeling in Generating Knowledge Components from Explanations. In *International Conference on Artificial Intelligence in Education*. Springer, 398–410.

[16] Korinn Ostrow and Neil Heffernan. 2014. Testing the multimedia principle in the real world: a comparison of video vs. Text feedback in authentic middle school math assignments. In *Educational Data Mining 2014*.

[17] Harold Pashler, Mark McDaniel, Doug Rohrer, and Robert Bjork. 2008. Learning styles: Concepts and evidence. *Psychological science in the public interest* 9, 3 (2008), 105–119.

[18] Thanaporn Patikorn and Neil T Heffernan. 2020a. Effectiveness of Crowd-Sourcing On-Demand Assistance from Teachers in Online Learning Platforms. In *Proceedings of the Seventh ACM Conference on Learning at Scale*. 115–124.

[19] Thanaporn Patikorn and Neil T. Heffernan. 2020b. Release of TeacherASSIST Dataset #1. (2020). DOI: http://dx.doi.org/10.17605/OSF.IO/EGP5F Accessed: 2020-05-15.

[20] Leena M Razzaq and Neil T Heffernan. 2009. To Tutor or Not to Tutor: That is the Question.. In *AIED*. 457–464.

[21] Paul Ruvolo, Jacob Whitehill, and Javier R Movellan. 2013. Exploiting commonality and interaction effects in crowdsourcing tasks using latent factor models. In *Neural Information Processing Systems. Workshop on Crowdsourcing: Theory, Algorithms and Applications*. Citeseer.

[22] Catharyn C Shelton and Leanna M Archambault. 2019. Who are online teacherpreneurs and what do they do? A survey of content creators on TeachersPayTeachers. com. *Journal of Research on Technology in Education* 51, 4 (2019), 398–414.

[23] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.

[24] John Stamper, Michael Eagle, Tiffany Barnes, and Marvin Croy. 2013. Experimental evaluation of automatic hint generation for a logic tutor. *International Journal of Artificial Intelligence in Education* 22, 1-2 (2013), 3–17.

[25] Daniel S Weld, Eytan Adar, Lydia B Chilton, Raphael Hoffmann, Eric Horvitz, Mitchell Koch, James A Landay, Christopher H Lin, and Mausam Mausam. 2012. Personalized Online Education-A Crowdsourcing Challenge.. In *HCOMP@ AAAI*. Citeseer.

[26] Jacob Whitehill and Margo Seltzer. 2017. A Crowdsourcing Approach to Collecting Tutorial Videos–Toward Personalized Learning-at-Scale. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. ACM, 157–160.

[27] Joseph Jay Williams, 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. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 379–388.

[28] Li Zhou. 2015. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326* (2015).