Effectiveness of Crowd-Sourcing On-Demand Assistance from Teachers in Online Learning Platforms

Thanaporn Patikorn
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts, USA
tpatikorn@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts, USA
nth@wpi.edu

ABSTRACT

It has been shown in multiple studies that expert-created on-demand assistance, such as hint messages, improves student learning in online learning environments. However, there are also evident that certain types of assistance may be detrimental to student learning. In addition, creating and maintaining on-demand assistance are hard and time-consuming. In 2017-2018 academic year, 132,738 distinct problems were assigned inside ASSISTments, but only 38,194 of those problems had on-demand assistance. In order to take on-demand assistance to scale, we needed a system that is able to gather new on-demand assistance and allows us to test and measure its effectiveness. Thus, we designed and deployed TeacherASSIST inside ASSISTments. TeacherASSIST allowed teachers to create on-demand assistance for any problems as they assigned those problems to their students. TeacherASSIST then redistributed on-demand assistance by one teacher to students outside of their classrooms. We found that teachers inside ASSISTments had created 40,292 new instances of assistance for 25,957 different problems in three years. There were 14 teachers who created more than 1,000 instances of on-demand assistance. We also conducted two large-scale randomized controlled experiments to investigate how on-demand assistance created by one teacher affected students outside of their classes. Students who received on-demand assistance for one problem resulted in significant statistical improvement on the next problem performance. The students' improvement in this experiment confirmed our hypothesis that crowd-sourced on-demand assistance was sufficient in quality to improve student learning, allowing us to take on-demand assistance to scale.

Author Keywords

Learning Management System; Crowd-Sourcing; Tutored Problem Solving; ASSISTments

CCS Concepts

•Information systems → Crowdsourcing; •Applied computing → Education;

L@S '20, August 12–14, 2020, Virtual Event, USA.
© 2020 Copyright is held by the author/owner(s).
ACM ISBN 978-1-4503-7951-9/20/08.
http://dx.doi.org/10.1145/3386527.3405912

INTRODUCTION

In recent years, the usage of digital learning media in K-12 classroom has grown exponentially. From the teacher and student perspectives, online learning platforms allow for new ways of learning that may otherwise be hard or impossible to do such as individualized mastery learning [5, 9]. Possibly one of the most important features of these systems is the ability to assist students as students work on their assignments. The most common type of assistants is answer feedback where the students know right away if their submitted answers are correct or not. In this work, we are interested in on-demand assistance. This type of assistance, sometimes called "tutoring," provides students with an option to request additional resources that would help them solve the problems, such as hint messages or complete explanations. The ability to provide students with additional guidance while they are outside of classrooms is especially valuable for homework assignments and distance learning such as during the COVID-19 pandemic. Since on-demand assistance is problem-specific, creating and maintaining on-demand assistance is hard and time-consuming. The cost of on-demand assistance scales with the number of problems in the system. For instance, out of 132,738 distinct problems assigned by all teachers inside ASSISTments in the 2017-2018 academic year, only 38,194 of them had on-demand assistance.

During a large-scale evaluation of the ASSISTments online learning platform [21], the intervention consisted of three components, supporting their 1) textbook work 2) skill builders (adaptive skill practice), and 3) teacher-created contents. First, related to the textbook work, we allowed each teacher to keep using their current textbook; we did the data entry to put the answers to the textbooks' questions into ASSISTments but we did not write hint message for them. When there is not a hint message, the student can try as many times as they want to answer the problem and they are only told if they are wright or wrong; if a student is totally stuck they can hit a button and be told the answer. We hypothesize that just seeing the answer will not help the student learn and hint messages will most likely be helpful. However, some studies have shown that certain hint messages may not always be helpful.

Skill builders are the second components of the [21] study. Teachers could choose to assign from over 200 skill builders that ranged in skills from adding whole numbers to quadratic equation solving. A skill builder gives students practice on a topic until they get three problems right in a row. All skill builders where built at WPI and every problem had hint message.

One of the teachers who participated in the large-scale evaluation inspired TeacherASSIST. Mr. Chris LeSiege, a teacher from Gorham, Maine, considered on-demand assistance to be of utmost important to his students' success. For the duration of the study, he added hint messages to most problems from his textbook. Unfortunately we did not anticipate that teachers would do this, there was no way for other teachers using the same textbook to review what Mr. LeSiege created and adopt it for their classrooms. At the time of the study, on-demand assistance was considered a component of the problem, thus, only the owner of the problem could edit or add on-demand assistance. For every problem he did not own, in particular the textbook problems created at WPI, Mr. LeSiege had to manually create his own versions of the problem and write the on-demand assistance on his version. His valiant efforts left us with two versions of the questions and more interestingly a larger question of how we should move forward as an educational platform. First, how can we better facilitate enthusiastic and diligent teachers like Mr. LeSiege. Second, since Mr. LeSiege spent a tremendous amount of time and effort to create hint messages, could we use them to help not only support his students but also all the other students working on the same problems? And how effective would they be for students outside of his classrooms?

Several studies shown that on-demand assistance created by experts increased learning outcomes [16, 19, 2, 24, 3, 10]. However, some studies suggested that on-demand assistance may not always be beneficial. For instance, assistance that is too detailed or provides too much information could result in less learning gain [11, 12, 23]. In addition, consistency of tones and pedagogical strategies could plays an important role in learning. [13], giving advice for those authoring textbooks, suggested that it is important to "establish a consistent standard." Lack of consistency, especially in difficult topics, could cause learners to miss important connections between terms and concepts across multiple learning materials. Thus its not obvious how effective crowd-sourced assistance would be?

The idea of using crowd-sourcing in K-12 education is not new. For example, Teachers Pay Teachers (teacherspayteachers.com) allows teachers to buy and sell their lesson plans and teaching materials. In fact, the 2019 American Instructional Resources Surveys showed that 56% of American Math teachers used resources from Teachers Pay Teachers [18]. Several educational researchers such as [27] and [25] also created proof-of-concept systems that crowd-sourced learning materials from MTurk workers and re-distributed them to MTurk workers/learners. They found that the learning gain from the best crowd-sourced materials was comparable to the learning gain from materials created by experienced instructors. Crowd-sourcing has also been used to accomplish other tasks in learning systems. For example, DALITE [4] and Ripple Learning (ripplelearning.org) crowdsourced instructions and resources generated by their peers. PeerWise crowdsourced multiple-choice questions from learners and re-distributed them to their peers [6]. Crowdsourcing had also been used to bring grading to scale such as [14], which is especially important in MOOCs. In fact, EdX, on of the most popular MOOC providers, is also a good example of how to use crowdsourcing to bring online MOOCs to scale.

While there are many examples of using crowdsourcing in educational platforms, to our knowledge, there has yet to be an example of a live system that crowd-sources contents from real teachers and redistribute this directly to real students without teacher intervention, AND that reliably improves student learning.

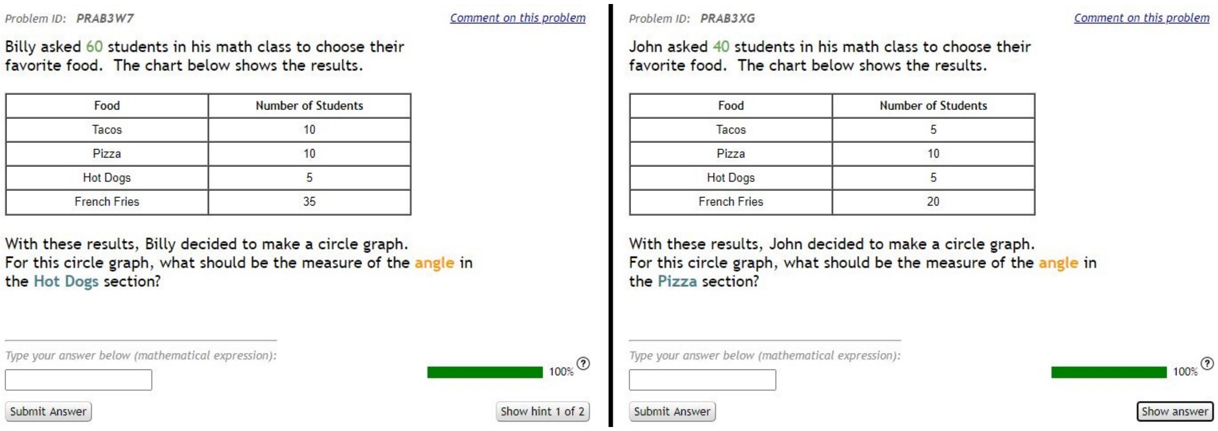
In this work, we designed and implemented a feature called TeacherASSIST inside ASSISTments to gather on-demand assistance by using crowd-sourcing. This component would later re-distributed crowd-sourced on-demand assistance to students outside of creators' classes. TeacherASSIST was created to answer our three research questions:

1. RQ1: "How could we design and implement a crowdsourcing system that allows teachers to quickly and conveniently create on-demand assistance for their students?"
2. RQ2: "How effective is such crowd-sourced assistance?"
3. RQ3: "Could we reproduce the same result if the same randomized controlled trial is run in a different academic term?"

BACKGROUND

In this work, we used ASSISTments an online tool used by teachers to support homework. ASSISTments (https://www.ASSISTments.org/) is a free online learning platform designed to empower teachers in their classrooms by automating laborious bookkeeping [9]. ASSISTments provides a library of problems, the majority of which is K-12 mathematics, that teachers can simple find, select, and assign to their students. ASSISTments provide immediate feedback as students work on their assignments and actionable reports to teachers. For every problem students receive instant correctness feedback, which tell the student whether the submitted answer is correct or not [8]. ASSISTments can also provide students with on-demand assistance, or "tutoring." Contrary to instant correctness feedback, on-demand assistance does not react to student answers. Rather, this type of assistance provides additional useful information and resources that help student solve the problem when requested (Figure 1). Both types of assistance have been shown to reliably improve student learning [19, 9, 20, 28, 15]. There are many types of on-demand assistance that had been shown to improve student learning such as step-by-step hints [9, 19], worked examples [7, 15], erroneous examples [15, 1], and providing the full solution to the problem [27, 25].

While many studies suggested that well-curated assistance improved student learning, there are also studies suggesting that some assistance may not be beneficial. A comprehensive literature review on the specificity of feedback and hint messages concluded that the literature is inconclusive on how specific feedback should be [23]. [12] showed that feedback with more information had a smaller effect on students' ability to correct



Problem ID: PRAB3W7

Billy asked 60 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	10
Pizza	10
Hot Dogs	5
French Fries	35

With these results, Billy decided to make a circle graph. For this circle graph, what should be the measure of the angle in the Hot Dogs section?

Type your answer below (mathematical expression):

Problem ID: PRAB3XG

John asked 40 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	5
Pizza	10
Hot Dogs	5
French Fries	20

With these results, John decided to make a circle graph. For this circle graph, what should be the measure of the angle in the Pizza section?

Type your answer below (mathematical expression):

Figure 1. (left) For this problem on-demand assistance (hint messages) is available. Thus, the student may click "Show hint 1 of 2" to request for hint messages. If no assistance is available (right), the student will only see "Show answer" which would mark the student as having given up on the problem.

their own errors than feedback with less information, such as providing only the correct answer. Another meta-analysis suggested that more-detailed feedback could result in worse learning outcome [11]. In addition, since most instances of on-demand assistance in studies were created by either experts in learning fields [16, 2, 24, 3] or by the instructors themselves [10], it would be dangerous to assume the same results for crowd-sourced on-demand assistance. In addition, since crowd-sourced assistance was created by neither experts nor their teachers, it is possible that the assistance could be of different tone or pedagogical strategies from those of the teachers or curricula. This inconsistency could reduce the effectiveness of learning materials and cause confusion [13].

There are several proof-of-concept studies on effectiveness of crowd-sourcing learning materials. For example, [25] crowd-sourced video lessons from MTurk workers and found that the learning gain from best crowd-sourced video was comparable to the learning gain from a popular video lesson from Khan Academy. Another system called AXIS [27] crowd-sourced explanations on how to solve a problem from MTurk workers. Then learners(other Mturkers were asked to revise and evaluate explanations as they solve problems. As learners work on problems, AXIS used machine learning to determine which explanations to present to to future learners. They found that explanations selected by AXIS were comparable to ones generated by experienced instructors, but all of this was done with Mturkers, not in authentic classrooms. To our knowledge, there is no live system that actively gets crowd-sourced assistance from teachers and directly redistribute them to students.

METHODOLOGY

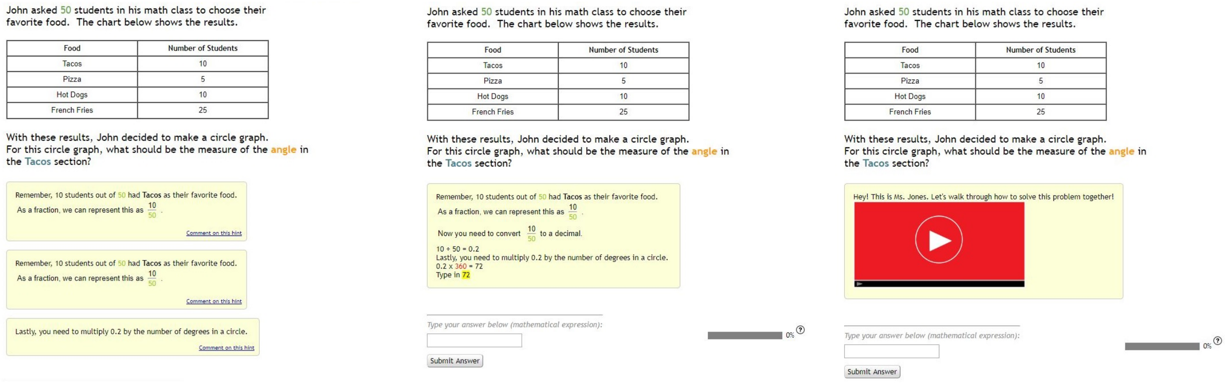
Before we designed and implemented the crowd-sourcing system for RQ1, we first investigated how to incentivize teachers to create on-demand assistance and designed an algorithm to distribute it. Then, we investigated the impact of crowd-sourced on-demand assistance on student learning. In this work, all the implementations, data collection, and analysis were done inside ASSISTments , our methodology is not platform-specific and should be applicable to other online learning platforms of similar characteristics and features.

Crowd-Sourcing On-Demand Assistance

For crowd-sourcing to be effective, we needed to obtain good quality on-demand assistance. The results of [27] shown that, given enough number of crowd-sourced on-demand assistance, we can obtain on-demand assistance of quality similar to one created by subject-matter experts. Thus, our goal was to design the system such that it is easy for teachers to create as much on-demand assistance as possible, as most users may not be motivated to contribute. However, as one of the main focus of ASSISTments and LMSs in general is to free teachers from laborious tasks, it is also important to not increase teachers' workload any more than needed. Thus, we collaborated with several teachers and investigated their normal everyday routines. The goal is to find the best approach to crowd-source on-demand assistance that are both convenient and beneficial to teachers' established routines for their classes and students.

The approach we took was first to create a component called "TeacherASSIST" inside ASSISTments. TeacherASSIST is a component allowed teachers to create on-demand assistance for their students as they taught the classes. Specifically, as teachers browsed through practice materials to assign to their students, they had an option to add their own on-demand assistance to each individual problem. This approach had many advantages. Firstly, teachers were incentivized to create on-demand assistance since it would directly benefit their students. Secondly, teachers were presented with the option to create on-demand assistance only for the problems they considered assigning to their students, so as not to overload them with too much to do. Lastly, the on-demand assistance was guaranteed to be of decent quality, as they belonged to the topics that teachers were currently teaching. Our implementation of TeacherASSIST was shown in Figure 4.

We then investigated what types of on-demand assistance should be supported. While we wanted to give teachers as much flexibility as possible, giving too many choices to the them could be detrimental and distracting [22]. We investigated the three types of on-demand assistance which were commonly available inside ASSISTments: hints, step-by-step problem-solving, and worked examples.



John asked 50 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	25

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

Remember, 10 students out of 50 had Tacos as their favorite food.
As a fraction, we can represent this as $\frac{10}{50}$.
Comment on this hint

Remember, 10 students out of 50 had Tacos as their favorite food.
As a fraction, we can represent this as $\frac{10}{50}$.
Comment on this hint

Lastly, you need to multiply 0.2 by the number of degrees in a circle.
Comment on this hint

John asked 50 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	25

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

Remember, 10 students out of 50 had Tacos as their favorite food.
As a fraction, we can represent this as $\frac{10}{50}$.
Now you need to convert $\frac{10}{50}$ to a decimal.
10 ÷ 50 = 0.2
Lastly, you need to multiply 0.2 by the number of degrees in a circle.
0.2 x 360 = 72
Type in 72

Type your answer below (mathematical expression):

John asked 50 students in his math class to choose their favorite food. The chart below shows the results.

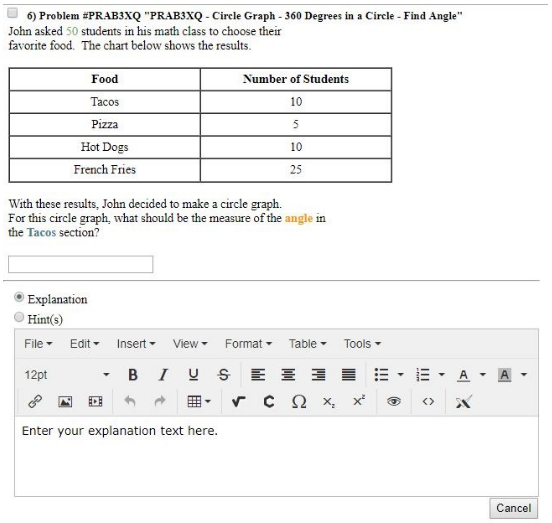
Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	25

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

Hey! This is Ms. Jones. Let's walk through how to solve this problem together!

Type your answer below (mathematical expression):

Figure 2. Examples of how the students see hints (left) and explanation (middle and right) in the ASSISTments tutor. Each yellow box in the left image represent a hint in the series. Explanations can be non-personal (middle) or personal (right).



□ 6) Problem #PRAB3XQ "PRAB3XQ - Circle Graph - 360 Degrees in a Circle - Find Angle"
John asked 50 students in his math class to choose their favorite food. The chart below shows the results.

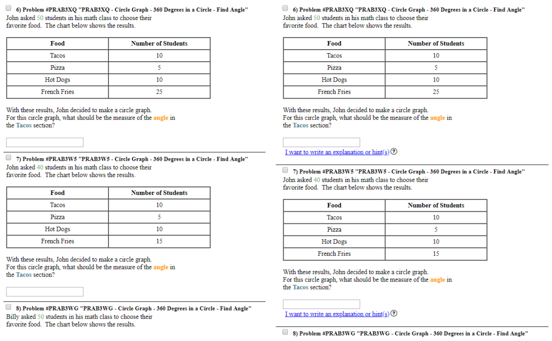
Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	25

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

● Explanation
○ Hint(s)

Enter your explanation text here.

Figure 3. Teachers can choose to create a set of hints or an explanation for any problems of their choice.



□ 6) Problem #PRAB3XQ "PRAB3XQ - Circle Graph - 360 Degrees in a Circle - Find Angle"
John asked 50 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	25

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

□ 7) Problem #PRAB3W5 "PRAB3W5 - Circle Graph - 360 Degrees in a Circle - Find Angle"
John asked 40 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	15

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

□ 8) Problem #PRAB3WG "PRAB3WG - Circle Graph - 360 Degrees in a Circle - Find Angle"
Billy asked 50 students in his math class to choose their favorite food. The chart below shows the results.

□ 6) Problem #PRAB3XQ "PRAB3XQ - Circle Graph - 360 Degrees in a Circle - Find Angle"
John asked 50 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	25

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

I want to write an explanation or hint(s)

□ 7) Problem #PRAB3W5 "PRAB3W5 - Circle Graph - 360 Degrees in a Circle - Find Angle"
John asked 40 students in his math class to choose their favorite food. The chart below shows the results.

Food	Number of Students
Tacos	10
Pizza	5
Hot Dogs	10
French Fries	15

With these results, John decided to make a circle graph.
For this circle graph, what should be the measure of the angle in the Tacos section?

I want to write an explanation or hint(s)

□ 8) Problem #PRAB3WG "PRAB3WG - Circle Graph - 360 Degrees in a Circle - Find Angle"

Figure 4. The interface where teachers find and assign a subset of problems inside a problem set without (left) and with (right) the option to create on-demand assistance for their students

1. Hints are a series of helpful messages that provide students with some information they need in order to solve a problem. Hints are usually given to students one at a time when requested. This means after students see each hint, they can attempt to solve the problem right away to show that they've learned the materials. Many systems take away a portion of partial credits if they request for hints.

2. Step-by-step problem solving or "scaffolding" problems is a type of on-demand assistance that breaks the original problems into smaller steps. The system will walk the students through each smaller step until the students reach the final "step" problem, which answers the original problem. This allows students with low prior knowledge or struggling students to learn how to solve complicated problems by filling their missing knowledge as they work on scaffolding problems. [19].

3. Worked examples provide full explanations on how to solve the similar problems, and sometimes the problem itself, from the beginning to the final answer. This type of on-demand assistance is analogous to teachers teaching students how to solve problems by demonstration.

We interviewed several teachers and educational researchers to find out the advantages and disadvantages of different types of on-demand assistance. In our final design, TeacherASSIST only allowed teachers to create hints and explanations, and not scaffolding problems. Creating scaffolding problems was complicated and time consuming, which is at odds with the narrative that teachers quickly create on-demand assistance as they assign problems to their students. In addition, even when the original problem is broken into smaller sub-problems, it is not uncommon for teachers to find struggling students stuck inside the "step" problems due to knowledge gaps.

The other two types of on-demand assistance, hints and explanations, have different advantages and disadvantages. On one hand, many teachers expressed that explanations were the easiest and fastest to create, as they had already been doing it while teaching. On the other hand, many educational researchers and teachers preferred hints to explanations since hints allowed students to demonstrate learning within a problem. However, teachers reported that it was harder to create hints in many topics without giving away the answer itself. It is also important to note that on-demand assistance is not limited to text; teachers were also allowed to include images, tables, and any types of formatting (Figure 3) and multimedia such as videos (Figure 2).

On-Demand Assistance Distribution

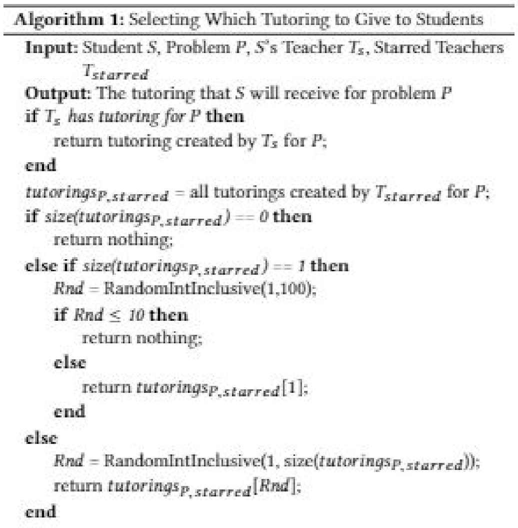
Before we distributed on-demand assistance, there were three major concerns we had to address. The first concern was privacy. While many teachers would not hesitate to create on-demand assistance for their own students, not as many felt comfortable sharing their on-demand assistance to students outside of their classes, especially if they included videos of themselves. Many teachers may not want to use on-demand assistance created by other teachers due to a different approach to solve the problems, which was the second concern. Lastly, as educational researchers, we wanted to be able to measure the quality of crowd-sourced on-demand assistance and to understand why each type of support suited different students through randomized controlled experiments.

In addition to the three concerns, there were three additional requirements that we considered to be most important. First, we needed to ensure that, if the teachers created on-demand assistance, their students must be guaranteed to receive them, regardless of what kinds of experiments were running and which other on-demand assistance is available. Second, since our main goal was to help students by providing them on-demand assistance as they are working on their assignments, it was important that such on-demand assistance be given out to as many students as possible. Third, we wanted to maintain the ability to conduct randomized control trials improve content as well as better on-demand assistance strategies.

As a result, we chose an approach similar to how new users in Wikipedia are promoted into confirmed and extended confirm users based on their activities [26]. For regular teachers, they can create any on-demand assistance for any problems. To address the first and second concern, such on-demand assistance will only be available to students in their own classes. Of those teachers, we searched for teachers who had regularly created on-demand assistance for their students and corrected any mistakes they found. With their consent, TeacherASSIST would re-distributed on-demand assistance created by starred teachers to students outside of their classrooms. This allowed us to scale-up on-demand assistance, addressing our second requirement. In order to satisfy the remaining concern and requirements, we came up with the distribution algorithm (Figure 5) that could run randomized controlled trials to determine the effectiveness of starred teachers' on-demand assistance.

Randomized Controlled Trials

TeacherASSIST was deployed in December 2017. We started promoting teachers to starred teachers in June 2018. Five teachers were promoted to starred teachers in 2018. Afterward, we started distributing starred teachers' on-demand assistance on October 10, 2018. The randomized controlled trial (named the "pilot experiment") started on the same date to answer RQ2. In 2019, we increased the number of starred teachers to nine and repeated the same randomized controlled trial (named "the repeated experiment") again to answer RQ3.



```
Algorithm 1: Selecting Which Tutoring to Give to Students
  Input: Student $S$, Problem $P$, $S$'s Teacher $T_s$, Starred Teachers
         $T_{starred}$
  Output: The tutoring that $S$ will receive for problem $P$
  if $T_s$ has tutoring for $P$ then
      return tutoring created by $T_s$ for $P$;
  end
  $tutorings_{P,starred}$ = all tutorings created by $T_{starred}$ for $P$;
  if $size(tutorings_{P,starred}) == 0$ then
      return nothing;
  else if $size(tutorings_{P,starred}) == 1$ then
      $Rnd$ = RandomIntInclusive(1,100);
      if $Rnd \leq 10$ then
          return nothing;
      else
          return $tutorings_{P,starred}[1]$;
      end
  else
      $Rnd$ = RandomIntInclusive(1, size($tutorings_{P,starred}$));
      return $tutorings_{P,starred}[Rnd]$;
  end
```

Figure 5. The algorithm we used for selecting which on-demand assistance should be given to a student for a given problem.

Specifically, the pilot experiment was conducted from August 9, 2018 to December 31, 2018 (corresponding to fall term of 2018). In this experiment, we compared crowd-sourced on-demand assistance (experimental condition) to simply giving the student the answer (control condition). For each problem with crowd-sourced on-demand assistance, the students were randomly assigned to one of the conditions at the problem-level. In other words, students could be in the control group for one problem, and in the experimental group for the next problem. We decided to use 9:1 as the ratio between the experimental condition and the control condition since we wanted to provide assistance to as many students as possible, and similar published works have shown similar on-demand assistance increases student learning. The repeated experiment was conducted and analyzed in the exact same manner as the pilot experiment, except it was conducted from January 1, 2019 to September 30, 2019 (corresponding to spring term and summer term of 2019).

When students worked on their assigned problems inside ASSISTments, they could see if there were on-demand assistance available before they requested it as seen in Figure 1. Specifically, if hint messages are available, students would see a button labeled "Show hint X of Y," where Y is the total number of hint messages available and X denotes which hint message will be given next. If no on-demand-assistance is available, the "Show answer" button will be displayed instead. Thus, we could not choose to analyze only students who requested for on-demand assistance since every student experienced the difference between condition, i.e. different buttons and corresponding partial credit costs, before receiving the treatment (i.e. requesting for on-demand assistance). Instead, we must first analyzed all students assigned to the control conditions and the experimental condition regardless of whether they actually requested for the assistance or not (we called this "intention-to-treat analysis"). After we determine that the button difference does not cause students in two conditions to behave significantly differently, we would then be able to analyze only students who request for assistance in the experimental condition or the answer in the control condition (we called this "treated analysis").

In the following section, we refer to the problems of where crowd-sourced on-demand assistance appeared as "RCT problems," and the math problems that the students worked on immediately after the RCT problems as "next problems." It is important to note that, for different students, the next problems were not guaranteed to be the same. In fact, for some RCT problems, the next problems may be in a different assignment, worked on a different day by the student. We will also use the term "ask for help" to refer to both students requesting on-demand assistance (experimental condition) and students requesting for the answer (control condition).

In this work, we only analyzed data where both RCT problem and the next problem come from to the same assignment.

In order to measure the quality of crowd-sourced on-demand assistance, we looked at 4 next-problem dependent measures.

1. "next problem correct first try": did the students answer the next problem correctly on their first try without using assistance or asking for the answer?
2. "next problem ask for help": did the students request for assistance or the answer during the next problem?
3. "next problem stop out": did the students give up solving the next problem?
4. "next problem attempt count": the number of attempts the student made during the next problem.

Our hypothesis was that the crowd-sourced on-demand assistance improved students learning. Students should be able to correctly answer the next problems more and ask for help less as they no longer need them. We did not expect a single problem-solving session to drastically change stop out rate or next problem attempt count. These two measures were included in the analysis to ensure that the differences between the correctness and help usages in the control condition and the experimental condition, if detected, were not caused by one of the conditions causing students to disproportionately give up on the next problems.

RESULTS

Overall Usage of TeacherASSIST

We investigated whether TeacherASSIST was able to incentivize teachers to create on-demand assistance. By the end of 2019-2020 academic year, three years after TeacherASSIST was deployed, we found that 146 different teachers had used TeacherASSIST to create 40,292 instances of on-demand assistance for 25,957 distinct problems across different curricula, 16,493 of which belong to our 9 starred teachers. Out of 146 teachers, 29 teachers had created more than 50 instances of assistance and 14 of those teachers created more than 1,000 instances of assistance over three years. To put the number in perspective, in 2017-2018 academic year, 132,738 distinct problems were assigned inside ASSISTments, only 38,194 of which had non-TeacherASSIST on-demand assistance. Of those problems, 27,094 more instances of on-demand assistance were created through TeacherASSIST, increasing the number of on-demand assistance by 70%.

Pilot Experiment

To measure how effective crowd-sourced on-demand assistance was (RQ2), we analyzed logged data of students who received on-demand assistance. We obtained problem log data from ASSISTments. For the duration of pilot experiment, there were 1,795 instances of on-demand assistance created for 1,787 unique problems. Out of instances of 1,795 on-demand assistance, 1,546 were explanations and 248 were hints. There were 142,010 problems solved in the randomized controlled trial, 128,153 of which received crowd-sourced teacher on-demand assistance and 13,857 of which only the answer was available. Our dataset is publicly available here [17].

Availability and Usages

Table 1 shows the availability and usages of teacher-created on-demand assistance and the crowd-sourced on-demand assistance. We found no significant difference between the percentage of students in the control condition and the experimental condition who answered the RCT problems correctly on their first try without asking for on-demand assistance ($p > 0.05$). Similarly, we found no significant difference between the percentage of students who requested crowd-sourced on-demand assistance (experimental) and students who requested for the answer (control) ($p > 0.05$).

Effects on Next Problems

To analyze the effects of crowd-sourced on-demand assistance on the next problems, we conducted the intention-to-treat (ITT) analysis. An intention-to-treat analysis is an analysis in which everyone who participated in the RCT is included in the analysis regardless of their scores, characteristics, and interaction with the intervention inside the RCT. Since our dataset was a problem-student level (i.e., a log of a student solving a problem), each observation was not independent (because one student solved multiple different problems and one problem was solved by multiple different students). Using t-test directly on the problem-student level would violate the independence observation assumption of t-test. Instead, we aggregated observations into 1) problem-level and 2) student-level, applied paired t-test on the aggregated observations, and reported the result of both aggregation methods for both the intention-to-treat and treated analysis. Since we performed multiple t-tests, we used the Benjamini–Hochberg procedure to obtain corrected p-values to reduce false positive.

In addition to the intention-to-treat analysis, we also looked at treated analysis. Treated analysis, in contrast with ITT analysis, only looks at participants who interact with the intervention or treatment. In our work, the treated analysis means that we would only look at students who asked for help while they worked on the RCT problem. The reason we also conducted the treated analysis was because a large majority of the students (67%) in both conditions were able to answer the RCT problems on their first try without requesting any on-demand

	number of problems solved	number of problems correctly solved on first try (percent)	number of problems where students requested for assistance or answer (percent)
teacher's own class	29049	19709 (67.84%)	4857 (16.72%)
control	13857	9377 (67.67%)	2271 (16.38%)
experimental	128153	86877 (67.79%)	20925 (16.32%)

Table 1. A table showing the availability and usages of teacher-created on-demand assistance and the crowd-sourced on-demand assistance.

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.65	0.66	-0.86	0.39	0.74
ask for help	0.17	0.16	0.86	0.39	0.74
stop out	0.03	0.03	0.48	0.63	0.74
attempt count	1.53	1.52	0.33	0.74	0.74

Table 2. Pilot Experiment: problem-level paired t-test intention-to-treat analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique problems = 1293

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.63	0.63	-1.35	0.18	0.23
ask for help	0.18	0.17	2.31	0.02	0.08
stop out	0.03	0.03	-0.81	0.42	0.42
attempt count	1.57	1.53	1.90	0.06	0.11

Table 3. Pilot Experiment: student-level paired t-test intention-to-treat analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique students = 4181

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.39	0.40	-0.80	0.42	0.42
ask for help	0.46	0.43	2.25	0.02	0.10
stop out	0.03	0.03	-0.85	0.40	0.42
attempt count	1.86	1.91	-1.04	0.30	0.42

Table 4. Pilot Experiment: problem-level paired t-test treated analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique problems = 620

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.39	0.41	-1.50	0.13	0.23
ask for help	0.45	0.41	3.39	<0.01	<0.01
stop out	0.03	0.04	-0.47	0.64	0.64
attempt count	1.91	1.85	1.35	0.18	0.23

Table 5. Pilot Experiment: student-level paired t-test treated analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique students = 1256

assistance. In addition, only a small portion of the students (16.7%) asked for help. This means the main difference between conditions (crowd-sourced on-demand assistance vs. answer) could be observed only on a small fraction of the students. Thus, in order to detect the effects in ITT analysis, the effects of the on-demand assistance must be very large to avoid being overshadowed by most of the samples that were not treated.

Intention-to-Treat Analysis

Table 2 and 3 shows the problem-level and student-level intention-to-treat analysis of the effect of crowd-sourced on-demand assistance using paired t-test. We found no significant difference between any next problem dependent measures using 5% false positive rate (alpha = 0.5) for Benjamini–Hochberg procedure. This is expected according to since Table 1 shows that, of all logged solved problems, more than 60% of the time students were able to solve the problems correctly on their first attempt without using on-demand assistance. In addition, students requested for on-demand assistance less than 20% of the time. In another word, a large majority of the students did not experience the difference between the control condition and the experimental condition.

Treated Analysis

Table 4 and Table 5 show the paired t-test of problem-level and student-level treated analysis of the effect crowd-sourced on-demand assistance. We found that, after applying Benjamini–Hochberg procedure, students who saw the on-demand assistance were less likely to request for more on-demand assistance in the next problem with statistical significance (corrected p-value < 0.01). This result can be interpret as either a positive or a negative effect of crowd-sourced on-demand assistance on learning. Students may either 1) learned enough to be able to solve the next problem, thus additional on-demand assistance was not needed, or 2) did not feel like on-demand assistance helps (e.g. of poor quality) and decided that requesting for any more on-demand assistance was not worth the partial credit cost. Using only the result data from the pilot experiment, we hypothesize that it was more like that crowd-sourced on-demand assistance had a positive impact on learning since, in addition to being well-supported by literature, while not statically significant, the percent of students in the experimental condition who answered their problem correctly on their first try is higher than that of the control, as well as with slightly lower attempt count.

Repeated Experiment

Using our data from the pilot study, we hypothesize that crowd-sourced on-demand assistance was of acceptable quality to improve student learning, causing them to answer more problems correctly while requiring less additional on-demand assistance. From January 1, 2019 to September 30, 2019, there were 232,248 problems solved in the randomized controlled trial, 208,987 of which received crowd-sourced teacher on-demand assistance and 23,261 of which only the answer was available. In said solved problems, 3,515 unique problems were solved with 3,698 distinct instances of on-demand assistance. Out of said on-demand assistance, 2,475 were explanations and 1,222 were hints. Similar to the pilot study, we found no significant difference between the percentage of students in the control condition and the experimental condition who answered the RCT problems correctly on their first try without asking for

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.62	0.63	-2.29	0.02	0.04
ask for help	0.20	0.19	3.65	<0.01	<0.01
stop out	0.03	0.02	1.47	0.14	0.19
attempt count	1.60	1.58	0.85	0.39	0.39

Table 6. Repeated Experiment: problem-level paired t-test intention-to-treat analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique problems = 2379

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.65	0.65	-1.59	0.11	0.15
ask for help	0.17	0.16	1.65	0.10	0.15
stop out	0.02	0.02	1.14	0.25	0.25
attempt count	1.60	1.55	2.86	<0.01	0.02

Table 7. Repeated Experiment: student-level paired t-test intention-to-treat analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique students = 6945

on-demand assistance (p > 0.05). We also found no significant difference between the percentage of students who requested for crowd-sourced on-demand assistance (experimental) and students who requested for the answer (control) (p > 0.05).

Intention-to-Treat and Treated Analysis of the Repeated Experiment

Interestingly, several tests from the intention-to-treat analysis are statistically significant even after using Benjamini–Hochberg with alpha = 0.05. Specifically, Table 6 shows that when aggregated on the problem level, students in the experimental condition were more likely to answer the next problems correctly on their first attempt as well as asking for less on-demand assistance on their next problem than students in the control condition (corrected p-value = 0.04 and <0.01, respectively). In addition, Table 6 showed that when aggregated on the student level, students in the experimental condition were more likely to have a smaller number of attempts than students in the control (corrected p-value = 0.02). Since inside ASSISTments, students were required to answer the problem correctly before they could move on to the next problem, a lower number of attempts meant students reached the correct answer faster on average, given there was no change in other dependent measures.

As for the treated analysis, the result aligned with the result of our pilot experiment. Table 8 and Table 9 show that the students in the experimental conditions asked for less on-demand assistance in the next problem (corrected p-value = 0.04 and correct p-value < 0.01 for problem-level and student-level aggregation, respectively). While not statistically significant, students in the experimental condition were more likely to answer the next problems correctly on their first attempt as well as asking for less on-demand assistance on their next problem than students in the control condition similar to the results we obtained from the pilot study and the intention-to-treat analysis.

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.36	0.37	-1.41	0.16	0.32
ask for help	0.49	0.47	2.54	0.01	0.04
stop out	0.03	0.03	0.01	1.00	1.00
attempt count	1.95	1.92	1.00	0.32	0.42

Table 8. Repeated Experiment: problem-level paired t-test treated analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique problems = 1312

next problem dependent var.	ctrl mean	exp. mean	t-stat	p-value	corrected p-value
correct first try	0.38	0.40	-1.86	0.06	0.08
ask for help	0.46	0.42	3.25	<0.01	<0.01
stop out	0.03	0.03	-0.68	0.50	0.50
attempt count	2.06	1.97	2.15	0.03	0.06

Table 9. Repeated Experiment: student-level paired t-test treated analysis on student next problem dependent variables using t-test and Benjamini–Hochberg. The number of unique students = 1955

CONCLUSION

In this work, we designed and implemented a mechanism that allows online learning platforms to crowd-source on-demand assistance from teachers. We developed this scheme in close collaboration with teachers and educational researchers to ensure that it is both convenient and beneficial to teachers, while remain open enough for researchers to conduct meaningful research.

To answer RQ1, we interviewed teachers and subject-matter experts to find out what are the features and requirements expected of on-demand assistance crowd-sourcing system, TeacherASSIST. Teachers wanted the system to improve student learning without overtaxing them and without additional work. Educational researchers wanted to be able to investigate the effectiveness of different kinds of on-demand assistance. Our ability to conduct RCTs for RQ2 and RQ3 shown that researchers can use TeacherASSIST to investigate the effectiveness of different kinds of on-demand assistance. While TeacherASSIST was designed and implemented inside ASSISTments, the core design and algorithm are applicable to other platforms that support on-demand assistance and content creation. Originally, only 38,194 of 132,738 distinct problems assigned inside ASSISTments in 2017-2018 academic year had on-demand assistance. By the end of 2019-2020 academic year, 27,094 instances of on-demand assistance were created for those problems through TeacherASSIST, starred teachers and otherwise. When we looked outside of the 2017-2018 dataset, we found a total of 40,292 instances of on-demand assistance across 25,957 distinct problems in different curricula, 16,493 of which belong to our 9 starred teachers. We also found that 14 teachers used TeacherASSIST heavily, creating more than 1,000 instances of assistance over three years.

To answer RQ2, we conducted the pilot RCT from August 9, 2018 to December 31, 2018. We found that students who requested the crowd-sourced on-demand assistance were reliably less likely to require additional assistance in the next problem. While the effect was small, it was expected since the experiment was conducted on the problem-level. Students who requested on-demand assistance were also more likely to correctly answer the next problem on their first attempt with lower overall average number of attempts, though it was not statistically significant.

To answer RQ3, we repeated the experiment we ran in RQ2 during the following academic term from January 1, 2019 to September 30, 2019. The results of the repeated experiment was in the same direction as RQ2, further confirm our hypothesis that crowd-sourced on-demand assistance is of high quality enough to improve student learning.

We concluded that we think the future of crowd-sourcing is bright. While there are several other crowd-sourcing applications in education such as [27] and [25], we are the first to crowd-source directly from active users (K-12 teachers) and redistributed crowd-sourced contents in a live environment. Our work serves as an evident that teachers are willing and able to create and improve contents of learning management systems, given that such contents are helpful to their students. We believed that a major part of this success was due to the fact that the designed of TeacherASSIST was heavily focused on teachers' need; TeacherASSIST was nicely integrated into teachers' routine and the on-demand assistance will directly benefit both their current and future students.

We also published the anonymized dataset from our large-scale randomized control trials. In this dataset, we included the data from both our pilot and the repeated experiments. All logged data (intention-to-treat) were included.

The code we used for analysis and datasets can be found here. https://doi.org/10.17605/OSF.IO/EGP5F

FUTURE WORK

In this work, our analysis is limited to next problem analysis. Ideally, we would like to measure student learning e.g. by using pre-test and post-test. However, since our randomized controlled trial was on an individual problem-level, it was impossible for us to have a proper pre-test and post-test. In order to solve that, we plan to design and run a different randomized controlled trial that would allow us to have some control over what the next and previous problems are using the problem set structure.

Alternatively, we could measure student learning by using more history and "future" information. For instance, we could compare the students history 10 problem before and after the RCT to get a better estimate of student learning. We would like to also look at the effects of on-demand assistance over multiples consecutive RCT problems as opposed to a single RCT problem. We expect this approach to have significantly bigger effect on student learning that what we have shown in this work.

In term of scalability, our method to aggregate on-demand assistance is currently naive. With better aggregation methods, we believe that the system would be able to select a better on-demand assistance, causing better improvement in student learning.

ACKNOWLEDGEMENT

We want to thank our funders that have helped support ASSISTments including NSF (e.g. 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 Schmidt Futures and another anonymous philanthropy. Most importantly, we thank the Office of Naval Research (N00014-18-1-2768).

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] 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.

[3] 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.

[4] Sameer Bhatnagar, Nathaniel Lasry, Michel Desmarais, and Elizabeth Charles. 2016. Dalite: Asynchronous peer instruction for moocs. In *European Conference on Technology Enhanced Learning*. Springer, 505–508.

[5] Albert Corbett, Megan McLaughlin, and K Christine Scarpinatto. 2000. Modeling student knowledge: Cognitive tutors in high school and college. *User modeling and user-adapted interaction* 10, 2-3 (2000), 81–108.

[6] Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. 2008. PeerWise: students sharing their multiple choice questions. In *Proceedings of the fourth international workshop on computing education research*. 51–58.

[7] 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).

[8] 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.

[9] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. 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.

[10] Kim Kelly. 2018. *A Set of Experiments Investigating Methods to Improve Student Learning Through Self-Regulated Learning*. Ph.D. Dissertation. Worcester Polytechnic Institute.

[11] Avraham N Kluger and Angelo DeNisi. 1998. Feedback interventions: Toward the understanding of a double-edged sword. *Current directions in psychological science* 7, 3 (1998), 67–72.

[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] Heng Luo, Anthony Robinson, and Jae-Young Park. 2014. Peer grading in a MOOC: Reliability, validity, and perceived effects. *Online Learning Journal* 18, 2 (2014).

[15] 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.

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

[18] Andrea Prado Tuma, Sy Doan, Rebecca Ann Lawrence, Daniella Henry, Julia H Kaufman, Claude Messan Setodji, David Matthew Grant, and Christopher J Young. 2020. American Instructional Resources Survey: 2019 Technical Documentation and Survey Results. (2020).

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

[20] Steven Ritter, John R Anderson, Kenneth R Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review* 14, 2 (2007), 249–255.

[21] Jeremy Roschelle, Mingyu Feng, Robert F Murphy, and Craig A Mason. 2016. Online mathematics homework increases student achievement. *AERA Open* 2, 4 (2016), 2332858416673968.

[22] Barry Schwartz. 2004. The paradox of choice: Why more is less. Ecco New York.

[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] 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.

[26] Wikipedia. 2020. Wikipedia:User access levels — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Wikipedia%3AUser%20access%20levels&oldid=960722670. (2020). [Online; accessed 11-June-2020].

[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] David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry* 17, 2 (1976), 89–100.