# How Common are Common Wrong Answers? Exploring Remediation at Scale

Anonymous Author(s)

## ABSTRACT

The process of synthesizing solutions for mathematical problems is cognitively complex. Students formulate and implement strategies to solve mathematical problems, develop solutions, and make connections between their learned concepts as they apply their reasoning skills to solve such problems. The gaps in student knowledge or shallowly-learned concepts may cause students to guess at answers or otherwise apply the wrong approach, resulting in errors in their solutions. Despite the complexity of the synthesis process in mathematics learning, teachers' knowledge and ability to anticipate areas of potential difficulty is essential and correlated with student learning outcomes. Preemptively identifying the common misconceptions in students that result in subsequent incorrect attempts can be arduous and unreliable, even for experienced teachers. This paper aims to help teachers identify the subsequent incorrect attempts that commonly occur when students are working on math problems such that they can address the underlying gaps in knowledge and common misconceptions through feedback. We report on a longitudinal analysis of historical data, from a computer-based learning platform, exploring the incorrect answers in the prior school years ('15-'20) that establish the commonality of wrong answers on two Open Educational Resources (OER) curricula–Illustrative Math (IM) and EngageNY (ENY) for grades 6, 7, and 8. We observe that incorrect answers are pervasive across 5 academic years despite changes in underlying student and teacher population. Building on our findings regarding the Common Wrong Answers (CWAs), we report on goals and task analysis that we leveraged in designing and developing a crowdsourcing platform for teachers to write Common Wrong Answer Feedback (CWAF) aimed are remediating the underlying cause of the CWAs. Finally, we report on an in vivo study by analyzing the effectiveness of CWAFs using two approaches; first, we use next-problem-correctness as a dependent measure after receiving CWAF in an intent-to-treat second, using next-attempt correctness as a dependent measure after receiving CWAF in a treated analysis. With the rise in popularity and usage of computer-based learning platforms, this paper explores the potential benefits of scalability in identifying CWAs and the subsequent usage of crowd-sourced CWAFs in enhancing the student learning experience through remediation.



*L@S '23, July 20–25, 2023, Copenhagen, Denmark*



ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

## CCS CONCEPTS

• **Computer Based Learning Platforms** → **Crowdsourcing**; • **Applied Computing** → *Education, Computer-Assisted Instruction.*

## KEYWORDS

Common Wrong Answers, Feedback Intervention Theory, Buggy Message, Engineering Feedback at Scale

**ACM Reference Format:**
Anonymous Author(s). 2023. How Common are Common Wrong Answers? Exploring Remediation at Scale. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (L@S '23).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The process of learning mathematics is cognitively complex. Solving a mathematics-based problem requires students to comprehend the problem requirements and demonstrate their knowledge and understanding of the topic. Students often deconstruct the problem task into smaller sub-tasks across multiple underlying concepts when synthesizing a solution. As part of the synthesis process, students practice various mathematical syntaxes, rules, and operations, reinforcing their knowledge and understanding of the underlying concepts that ultimately help students learn and develop their understanding of the main problem task. While learning and synthesis appear intuitive and easily understandable, analyzing such processes can be particularly challenging; the individual steps to reach a solution are not obvious because synthesis is a process that is intrinsic to the learner. Students can leverage their intrinsic cognitive ability to take various approaches to synthesize a solution. The approaches can differ in various ways, e.g., the complexity of the broken-down sub-task and the order in which the sub-tasks are solved.

Despite the variation in the approach, students require a fundamental understanding of the mathematical processes to solve the problem; however, gaps in student knowledge or misconception, errors on one or more steps in solving a problem due to misconception or "slip"[12] can lead to a variety of incorrect responses. Alternatively, gaps in student knowledge and shallow learning concepts can cause students to guess at answers or apply the wrong problem-solving approach, resulting in an entirely different set of incorrect answers [7]. Regardless of the cause, without directed feedback on how to resolve errors experienced during problem-solving, the errors may impede a student's learning progress. Understanding the common errors that students experience as they interact with mathematics-based problems is critical for guiding the design of effective instructional practices to help students learn correct mathematical processes and problem-solving strategies [28].

The process of diagnosis and examination of "Common Wrong Answers" (CWAs) is critical to understanding learning processes in the context of mathematics. CWAs may be utilized to develop better

educational technologies that, in conjunction with teachers, can address the needs of individual students–educational technologies often referenced as Computer Based Learning Platform (CBLP), Online Learning Platforms (OLP), or Intelligent Tutoring Systems (ITS).

In this paper, we leverage historical data on a CBLP in the analysis of CWAs on Open Educational Resource (OER) curricula: Illustrative Math (IM) and EngageNY (ENY) for students in grades 6, 7, and 8 across 5 school years. Through the analysis, we explore the commonality of CWA across multiple academic years with shifts in the underlying student and teacher population working on the problems. We then extend our analysis by conducting goals and task analysis in an engineering crowdsourcing platform that teachers can use to develop Common Wrong Answer Feedback (CWAFs). CWAFs aim to address student misconceptions and gaps in knowledge by providing instructional guidance that nudges the students towards the solution while addressing the error in their approach whenever applicable. Finally, we conduct a within-subject-problem-level randomization exploring the efficacy of CWAFs at scale by using next-problem correctness in a treated analysis [1].

### 1.1 Research Questions

Toward the exploration of "How common are CWAs?" and "Can we remediate them?", the paper addresses the following main research questions:

(1) Do students commonly make similar errors when working on math problems?
(2) What fundamental goals and tasks must a crowdsourcing platform provide when facilitating the generation of CWAF?
(3) Does the remediation of CWAs with CWAFs lead to better learning outcomes?

## 2 BACKGROUND

### 2.1 Common Wrong Answers

Common Wrong Answers (CWAs) are common mistakes or errors that typically arise from a buggy rule, a common misconception about the topic, or a lack of knowledge among the students. Several prior research in the domain of cognitive science and mathematics learning have investigated the common errors made by students during the process of solving a mathematics-based problem [7–9, 29, 43, 44].

Some of the prior works on this [11, 36] have explored the rectification of these common errors in students understanding through instruction. Brown et. al [7] in their prior work analyzed common student mistakes when solving multi-digit subtraction problems and used their analyses to develop a diagnostic model that detects and explains these errors. Brown et. al further [8] introduced the "generative theory of bugs", which is a set of formal principles to explain the known/common errors in a procedural skill. Other studies like Sison et. al [37] present student modeling approaches in identifying the common errors or bugs in student works. In their work, they also talked about the importance of identifying a "bug library", which is the collection of the most common misconceptions or errors made by a population of students. However constructing these libraries is a challenging task, as these misconceptions vary based on the population of students and different groups of students may exhibit different types of misconceptions when synthesizing a solution for mathematical problems.

In addition to the principles of learning theory and cognitive skill acquisition, various prior research [36] have also explored the potential of algorithmically identifying the common misconceptions of students, to further rectify the incorrect and buggy processes in students' work. Selent et al. [36] explored the use of machine learning methods to predict CWAs and their causes. Further, they explored the effectiveness of suggesting buggy messages when a student makes common mistakes. With the use of these buggy messages to rectify common errors in students' work, they measured the reduction of help-seeking behavior in an online learning platform.

### 2.2 Feedback Intervention

Feedback is one of the major factors influencing learning outcomes and achievement. However, the impact of feedback depends on the type and way of delivery. Prior research on Feedback Interventions (FI) through conducted meta-analyses has produced mixed results on their effectiveness on student performance [1, 2, 16, 20, 23, 33, 38, 39]. The results from these works have led to further research toward the exploration of the nuances of FI, resulting in the development of Feedback Intervention Theory (FIT) [20]. FIT operates under the assumption that FIs aim to catch the recipient's attention across 3 hierarchically organized levels: task learning, task motivation, and meta-task. While there are concerns regarding the general effectiveness of FIs it is much less of a concern in an educational context. Hattie, 1999(c.f.[16]) reported on a synthesis of over 500 meta-analyses exploring the effect of schooling on students where FIs[1, 2, 23, 33, 38, 39] were among the top 10 highest influences on the student achievement–highlighting the effectiveness of FIs in learning.

Effective Feedback can help learners track their progress, validate the students' effort, reinforce their progress, and impact their reactions and behavior when working on activities[10, 15, 45]. While feedback is indeed crucial to the student's learning experience, the quality of the feedback varies greatly. Studies, such as [42] present the importance of student perception in the effectiveness of feedback. In their work, they report that detailed constructive feedback from instructors were found to be the most beneficial, and if the feedback was too vague, or did not have enough content, the usefulness of the feedback would wane. Studies, such as [22], discuss that providing feedback in an online setting is an art and that there are various best practices including generating positive feedback and/or balanced feedback. In this paper, we focus on the exploration of tailored feedback in the remediation of common bugs in students' work; as such, we use the Hattie et al.(c.f. 2007) [17] conceptualization of feedback [2]. Hattie et al., 2007[17] expanded upon the generalized FIT model and proposed a theoretical model aiming

[1]The data and code used in this paper is shared through open-science practices at BLINDED-URL

[2]*[17] Feedback is conceptualized as information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding. A teacher or parent can provide corrective information, a peer can provide an alternative*

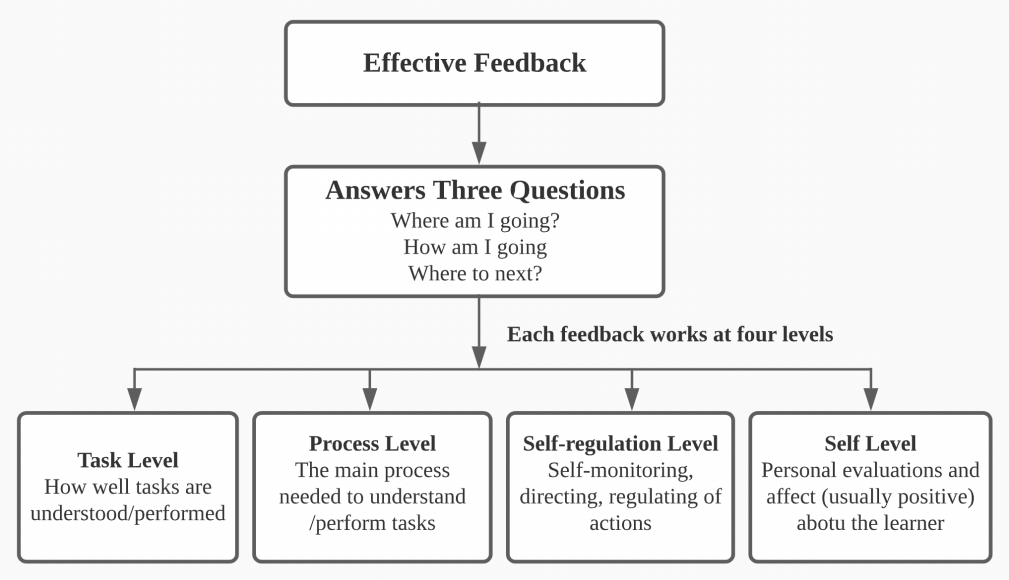


**Figure 1: A model of feedback for enhanced learning, taken from Hattie et. al [17]**

to reduce the discrepancy between the current and desired understanding of learners in an educational context. Figure 1 presents the theoretical feedback model proposed by Hattie et. al [17] for enhanced learning. The theoretical model posits that the feedback must answer three major questions: (1)What are the goals? (2)What progress is being made toward the goal? (3)What activities need to be undertaken to make better progress? The FIs address these 3 major questions by operating across operate on 4 levels of instruction: (a) task level, (b) process level, (c) self-regulation level, and (d) self-level. Hence, effective feedback needs to incorporate the following characteristics: recognize if the task requirement is understood, exhibit the correct processes required to complete the task, include instructions that direct the learner towards the following productive actions, and include evaluation and affect(usually positive) to personalize the instruction.

## 2.3 Common Wrong Answer Feedback

Remediation of common errors in students' work has been a focus of several prior research[26, 27]. Vanlehn et. al [41] in their study, observed the interaction between expert human tutors and physics students, and they study the effect of tutor explanations to remedy student errors. In their study, they find that only some explanations are associated with improved learning and that the effectiveness of the feedback varied with the content and question. In addition, short and concise explanations to remedy errors were observed to be more effective compared to longer and more elaborate explanations. Other studies [35] have identified the inability of guided instructions in the remediation of errors that arise from misconceptions from previously learned skills, suggesting that deeply ingrained misconceptions and bugs may be more difficult to rectify over time.

*strategy, a book can provide information to clarify ideas, a parent can provide encouragement, and a learner can look up the answer to evaluate the correctness of a response. Feedback thus is a "consequence" of performance.*

Studies like [14, 21, 34] introduced the use of error analysis methods as a step towards understanding students' ability to identify and explain errors in given problems. They presented students with erroneous examples and asked them to detect and explain the error in the given examples. Rushton et al. [34], reported on the approach of error analysis leading to better knowledge retention over the traditional methods of learning mathematics.

## 2.4 Crowdsourcing Instruction

Crowdsourcing has become an increasingly popular method in K-12 education for gathering feedback on instructional materials [13]. With the aid of various authoring tools, teachers and educators are able to create and distribute more representative educational content. Many online learning platforms and tools [4, 18] support crowdsourcing of instruction and teacher-authored content. Research in the educational context has shown that crowdsourcing can improve the online learning experience by providing students with on-demand teacher support, tutoring, hints, and explanations[19, 24, 30, 32]. Some other studies[3, 5] have also explored the use of crowdsourcing to collect teacher-given scores and feedback messages for students' open-ended math answers. Prior research around this [31, 32] have demonstrated the effectiveness of crowdsourcing instruction and tutoring content in online learning platforms toward enhancing the quality of instructional materials and the learning experience for students. As such in this study, we intend to crowdsource CWAFs through the development of a crowdsourcing platform for teachers toward the remediation of CWAs.

# 3 EXPLORING COMMON WRONG ANSWERS

The exploration of the commonality of CWAs was conducted by examining data from students in grades 6, 7, and 8 who worked on problems in two commonly used mathematics curricula: Illustrative

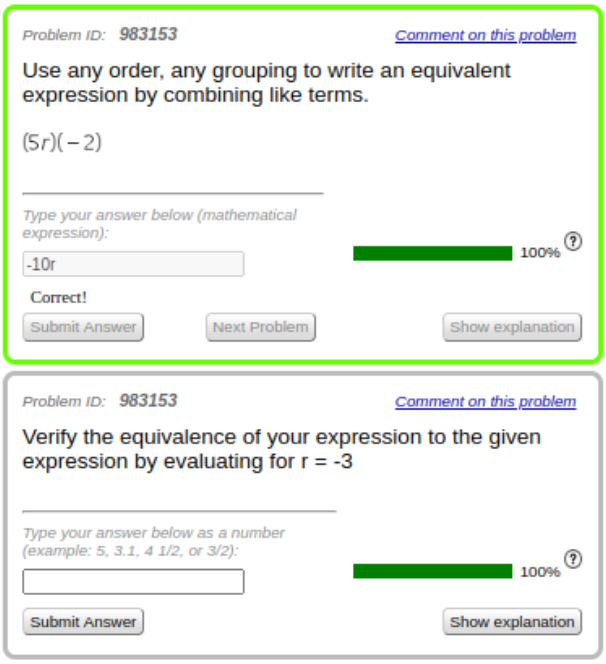


**Figure 2: An example of two consecutive problems in a problem set that the student is working on that has the same set of Common Core Standards associated with it.**

Mathematics (IM) and EngageNY (ENY) over a five-year period from '15-'16 to '19-'20. The students' data were collected from a CBLP [BLINDED FOR REVIEW]. A summary of the total number of problems the students worked on across the 5 school years from '15-'16 to '19-'20 is presented in table 1–the problems were considered eligible for the count if they were worked on by more than 20 students in at least one of the five school years. We observe that ENY on average is used more often than IM and on average teachers have used the content for grade 7 ENY the most across the 5 academic years.

Within the CBLP in this study, students typically work on a set of assignments with multiple problems that may or may not be associated with the same set of skills (Common Core Standards). Figure 3 presents an example of two consecutive problems from the ENY curriculum within the same set of Common Core Standards. These two problems are associated with the same skill set – the first question is asking to simplify the equation and the second question is asking to verify the results. These problems have a higher likelihood of knowledge transfer than problems coming from different common core standards.

To investigate the frequency of incorrect responses, we analyzed the first incorrect attempt made by each student while working on the problems. Using this approach, we generated the top 3 most common incorrect answers (CWAs) for each problem. We added an additional condition to help increase the reliability of the CWAs by only analyzing the problems that had been worked on by at least 20 students during the school year and more than 10 students had made the most common wrong answer. We observed that 1,045 problems had CWAs across at least two academic years. An example of CWAs across academic years is provided in table 2. We observe that the common wrong answers for the problem shown in 3 meet the threshold of commonality in 4 out of the 5 academic years with the first CWA replicating in each year, however, the second and third CWAs did fluctuate with some commonality across years where the ranks of some of the CWAs were swapped whereas certain academic years had completely new CWAs were observed. Further, we can observe that the number of students is decreasing across school years, this decline can be attributed to a version upgrade to the CBLP used in our analysis–teachers began migrating to the newer version during the '18-'19 academic year. While the reduction of the number of students did decrease the total number of students available for our exploration of CWAs in the later academic years it doesn't prevent us from demonstrating their prevalence as the same CWAs repeated despite the changes in the student and teacher population working on the problems. A more extensive example of the CWAs from our analysis is in the supplementary materials accompanying this paper.

From the exploratory analysis exploring the occurrence of CWAs we observed that CWAs repeat across school years. Upon deeper analysis of the problems with CWAs we observed that the majority of the problems belonged to Practice Problems for the lesson component that are designed to help students learn the content. In the next section, we elaborate on an iterative process of goals and task analysis that informed the design and development of a crowdsourcing tool for teachers to write CWAFs aimed at remediating the gaps in student knowledge that resulted in the CWAs.

## 4 TASK ABSTRACTION

In this section, we elaborate on our design and development process of a crowdsourcing tool. In order to understand the requirements of a crowdsourcing tool, we interacted with various experienced teachers, teacher trainers, domain experts, and researchers exploring crowdsourcing tools. Overall our analysis can be divided into two parts. We began with a goals analysis where we developed a hierarchy of goals the tool needs to facilitate. These goals were deconstructed into sub-goals that directly correlate with teacher needs. We leverage the concrete set of subgoals to enumerate visualization components in the crowdsourcing tool that directly facilitates the teachers' needs. We utilize the "Nested Model for Visualization" ( c.f., [25]) to develop a better understanding of the fundamental goals and tasks of a crowdsourcing platform. We utilize the design language proposed by Munzner, 2009 [25] in the design and development of our tool to provide some fundamental guardrails to similar projects exploring crowdsourcing in the future.

First, we establish the high-level goals and sub-goals a tool needs to facilitate. Upon validating the goals and subgoals with end users and domain experts, we conducted task abstraction to define low-level tasks such as browsing, exploring, and identifying from the Brehmer and Munzner topology that designers and developers can utilize in developing the tool. While the crowdsourcing tool we developed in this paper doesn't contain elaborate visualization components traditionally associated with most visualization projects, The "Nested Model for Visualization" and Brehmer and Muzner's topology (c.f., [6]) can be highly effective at identifying goals and

**Table 1: Summary of Total Problems and Problems with CWAs. The problems with CWAs met our threshold of more than 20 students working on the problem in at least two consecutive years and more than 5 students making the same common wrong answer in each year.**

| | Engage NY | | Illustrative Math | |
|---|---|---|---|---|
| *Academic Level* | Total Problems | Problems with CWAs | Total Problems | Problems with CWAs |
| Grade 6 | 1351 | 210 | 2082 | 254 |
| Grade 7 | 1845 | 511 | 2088 | 518 |
| Grade 8 | 1076 | 92 | 1475 | 267 |

**Table 2: Common Wrong Answer by Student Count**

| | | | First CWA | | Second CWA | | Third CWA | |
|---|---|---|---|---|---|---|---|---|
| Number of Students | Incorrect Count | Correct Answer | Answer | Count | Answer | Count | Answer | Count |
| 214 | 62 | 30 | -30 | 42 | 5 | 5 | 13 | 2 |
| 354 | 75 | 30 | -30 | 44 | -17 | 3 | -13 | 5 |
| 332 | 98 | 30 | -30 | 71 | -17 | 5 | 0 | 3 |
| 243 | 63 | 30 | -30 | 38 | -15 | 4 | -17 | 4 |

tasks that can be leveraged in augmenting the ability of the teachers in formulating effective feedback–the overarching goal of our crowdsourcing tool. As such, we leverage design techniques popular in the HCI(Human Computer Interaction) domain to conduct task abstraction and generalize a crowdsourcing tool's development process. We went through multiple iterations of goals and task analyses to further refine our findings and report on the final result in this section.

## 4.1 Goal Analysis

Table 3 lists the goals and sub-goals resulting from our analysis. The overarching goal of the tool is to augment teacher ability in gaining insight into the various process the students might have taken that resulted in a "bug" during the synthesis of a solution that resulted in the CWAs. While the underlying mechanism that resulted in the CWAs is unknown, we aim to leverage teacher experience and intuition to discern the underlying cause and generate appropriate feedback intervention to help remedy the cause. We identified 3 distinct goals a crowdsourcing tool needs to facilitate. The first two goals, G1, and G2, directly address teacher needs in substantiating the CWAs and providing contextual insight to help teachers formulate effective feedback. Goal 1 helps teachers understand the general student performance on the problem, provide evidence towards the commonality of the response, and identify the problems within a set of problems where students struggle the most, i.e., most likely problems within a set of problems where gaps in student knowledge will impact their performance the most. The intent of goal 2 is to provide contextual information that can augment teacher ability when analyzing the CWAs and their potential causes by providing contextual information. Additionally, information on prior problems related to the same skill component can provide scaffolding that teachers can leverage in contextualizing the problems and converging on a smaller subset of potential causes for the CWAs. While the primary objective of the tool is to facilitate the generation of CWAFs, both the teachers and domain experts on multiple occasions emphasized the importance of goal 3 in dictating the quality of the feedback through collaboration and validation from peers.

## 4.2 Task Analysis

For each sub-goal presented in table 3 we generated a list of low-level sub-tasks designed to help teachers (a) look up other problems within the problem set, (b) explore various knowledge components the students struggled with while working on the problems, (c) identify the potential causes of the CWAs, and (d) produce feedback that can effectively help remediate gaps in student knowledge that resulted in the CWAs. These sub-tasks are related to the abstract visualization task from Brehmer, and Munzner's topology [].

Table 4 presents the high-level task that can help inform the design and development of features within the crowdsourcing tool that can help facilitate one or more sub-goals which in turn, in conjunction with other tasks, can help achieve the main goals of crowdsourcing. While the tasks can be further decomposed into auxiliary sub-tasks that can be specific to the objective of the project, as such, we only report on the high-level task analysis. We find the tasks to be self-explanatory; as such, we refrain from elaborating upon the tasks at length in the text to avoid redundancy and preserve space. Furthermore, it is important to note that this list is not intended to be an exhaustive list that describes what constitutes an effective crowdsourcing tool but rather a reference for others to what we observed to be useful from our interaction with teachers and other stakeholders during the design and development of the tool.

# 5 CROWDSOURCING COMMON WRONG ANSWER FEEDBACK

In this section, we briefly describe our implementation of the crowdsourcing tool guided by the goals and task analysis described in the

**Table 3: Fundamental goals of a crowdsourcing tool.**

| Generic Goals | | |
|---|---|---|
| **G1** | **Substantiate the Common Wrong Answer** | |
| | a | Analyze general student performance on the problem. |
| | b | Validate the common wrong answer. |
| **G2** | **Contextualize the Common Wrong Answer** | |
| | a | Identify problems where students struggle the most. |
| | b | Identify the underlying mechanism for the common wrong answer. |
| **G3** | **Facilitate Collaboration and Support.** | |
| | a | Facilitate alternative perspectives to edify teachers' understanding of the problem requirements. |
| | b | Facilitate collaboration and validation through peers support. |

**Table 4: Task analysis of each sub-goal.**

| Tasks | |
|---|---|
| **G1. a. Analyze general student performance on the problem.** | |
| T1 | Identify problem properties, e.g., general difficulty, problem type, and answer. |
| T2 | Identify student performance on a problem, e.g., total students, percent correct. |
| **G1. b. Validate the common wrong answer.** | |
| T3 | Examine the CWAs, e.g., incorrect answer, frequency of CWAs. |
| T4 | Verify the CWAs is caused by mathematical error and not due to underlying bugs in the system. |
| **G2. a. Identify problems where students struggle the most.** | |
| T5 | Examine the problems within a problem set where students perform poorly. |
| T6 | Identify the knowledge components required to do well on the problem set. |
| T7 | Infer the amount of effort and attention required to solve the problem. |
| **G2. b. Identify the underlying mechanism for the common wrong answer.** | |
| T8 | Identify if the cause of the CWAs, e.g., misconception, gaps in knowledge, trick question, slip, or guess. |
| T9 | Examine if the CWAs is influenced by a prior problem or if the problem will cause CWAs in the future. |
| **G3. a. Facilitate alternative perspectives to edify teachers' understanding of the problem requirements.** | |
| T10 | Identify opportunities for the teacher to analyze the CWAs from multiple perspectives, e.g., feedback for high-knowledge students, feedback to teachers when their students struggle with the problem. |
| **G3. b. Facilitate collaboration and validation through peer support.** | |
| T11 | Facilitate peer collaboration, e.g., synchronous and asynchronous pair work. |
| T12 | Enable teachers to review each other's feedback. |

prior section. In order to facilitate the fundamental goals described in table 3 we designed a new tool within the CBLPs. The tool allows teachers to identify relevant CWAs, gain contextual insight into the problems associated with the CWAs, and facilitates peer collaboration to help further improve the quality of the CWAs.

Figure 3 shows the teacher view of a teacher working on a problem set in IM curricula for grade 7, unit 8, lesson 8 based on the common core standard for teaching "Probability and Sampling". As shown in the figure, a teacher has analyzed the first CWA for the problem and provided appropriate feedback addressing the student's misconception. The teacher can substantiate the CWAs by examining the number of students that have worked on the problem, the percentage of students who answered it incorrectly, identifying the top 3 CWAs, and the percentage of students who made the CWAs among students who answered it incorrectly. Besides examining the validity of the CWAs the teacher can also explore the other problems in the problem set and their respective CWAs to gain insight into how students working on the problems have historically struggled. We posit that this insight, combined with the ability to collaborate with peers and review each others' work, can facilitate the generation of effective CWAFs.

As the primary objective of this paper is to examine CWAs and investigate the fidelity of CWAFs in remediating the underlying causes of the CWAs, we hired 24 teachers, teaching IM and ENY in middle school, in two batches to help write CWAFs for problems in IM and ENY for grade 7. As teachers in our initial analysis of CWAs primarily used the "Practice Problems" as opposed to "Exit Tickets" in ENY and 'Student Facing Task" and "Cool Down" in IM, respectively, the 24 teachers were asked to write CWAFs for problems in the "Practice Problems" section. Furthermore, as the generated feedback is intended to be used by middle school students, the teachers were given an initial introductory training by domain experts regarding the structure of the feedback and the use of certain mathematical terms to ensure that the feedback remains consistent with the mathematical terms used in each curriculum. The domain experts also provided feedback on the CWAFs once the teachers started writing feedback. Finally, the domain experts also functioned as moderators to ensure consistency in the quality of

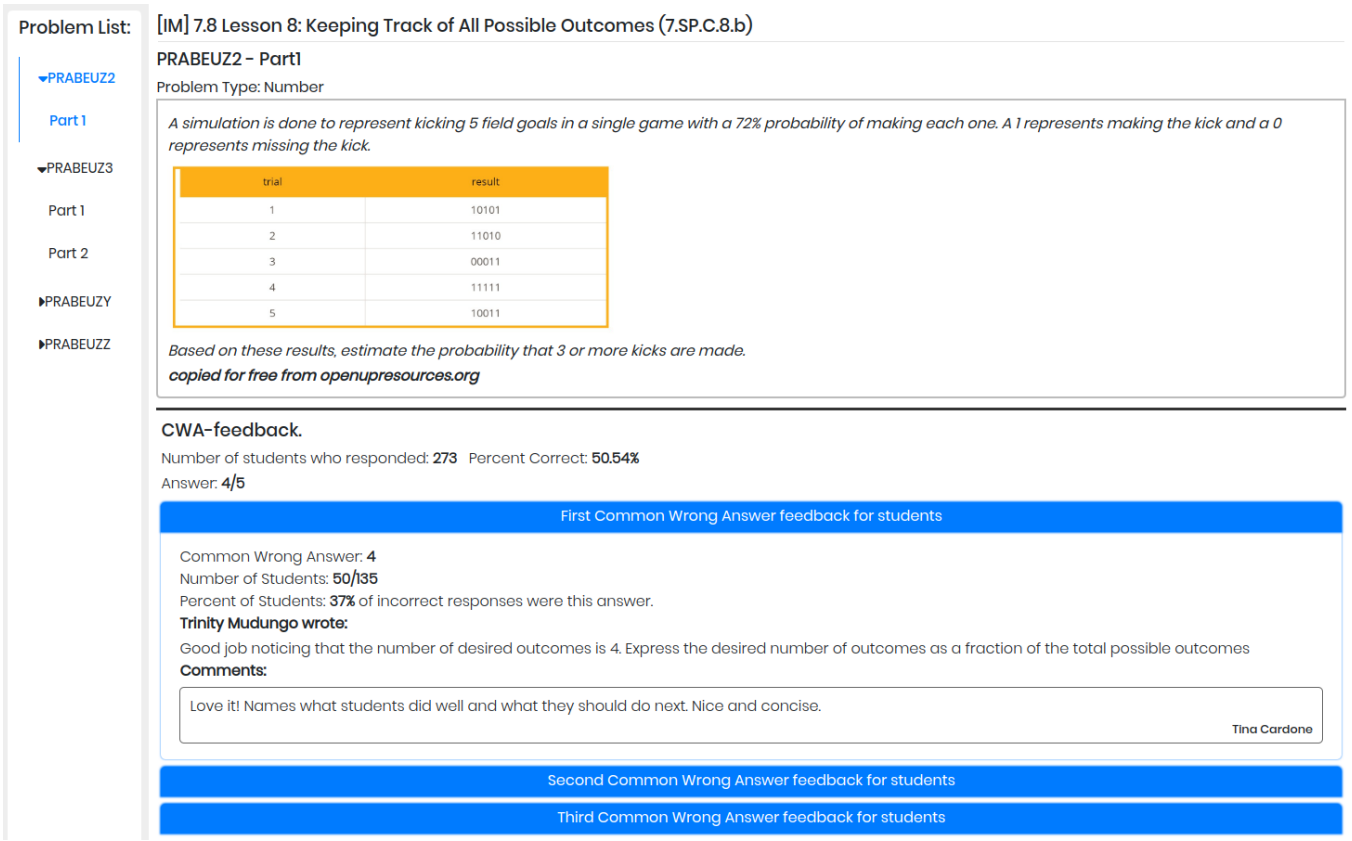


**Figure 3: Teacher perspective, visualization of a problem from Illustrative Math curricula with Common Core standard 7.SP.C.8.b where a teacher has written feedback and a peer/moderator has reviewed it as well.**

the CWAFs, and they were ultimately responsible for approving the feedback to be eligible for use with students. The extra precaution was taken to avoid any possibility of accidentally exposing students to any harmful content.

In the next section, we describe a within student problem level randomized control trial that we conducted to examine the efficacy of CWAFs at scale.

## 6 IMPLEMENTING COMMON WRONG ANSWER FEEDBACK

The crowdsourced CWAFs collected from the teachers were integrated into the CBLP upon approval from the moderators. The initial implementation of the first batch of the CWAs was conducted in April of '22. Over time and across multiple iterations, the CWAFs, crowdsourced from teachers for 1,660 problems, have been integrated into the system and are actively provided to students.

### 6.1 Experimental Design

Students are randomized to either a control or a treatment condition during the assignment, i.e., students are assigned to business as usual(no CWAF) or treatment(CWAFs) when they begin a problem. Ideally, randomizing students once they make a CWA would be more optimal; however, the process of triggering a server request that randomizes students once they enter a CWA can take away from the learning experience of the student and can ultimately hamper their perception and usage of the platform itself as such we randomize at the assignment and analyze the effectiveness of CWAFs on the treated group. The students are randomized in a 90:10 split where there is a 90% chance that the student will be assigned to a treatment and a 10% chance of being assigned to control. The 90:10 split was chosen to ensure that most, if not all, of the students, had multiple opportunities to access the CWAFs and learn from the feedback. At scale, a 90:10 randomization should still have enough power to help explore the effectiveness of our treatment without impeding students' access to learning opportunities.

### 6.2 Dataset

Since the initial implementation of the first batch in April '22, CWAFs have been randomized across 20,044 students working on 1,387 problems in ENY and IM a total of 623,857 times; students were assigned 560,897 times to treatment and 62960 times to control. While the students were assigned to treatment or control, they only received CWAFs if their attempt was one of the top 3 CWAs for the problem. As such, we dropped the students who did not attempt to answer the problem with a CWAs at any point while working on the problem. After dropping the students who did not

make any attempts that identified as a CWA for both control and treatment, we have 14,341 unique students who were randomized and made at least one CWA when working across 1,018 problems. With this, we have 94,765 students in treatment and 10,817 in control. While this data is for students working on problems within the same problem set, different problems in a single problem set can have different sets of common core standards associated with different skills. As such, we filter the treated students to examine the effectiveness of CWAFs by only analyzing the problems where both the intervention and the next problem had the same common core standards. This additional filtering requirement reduced the number of distinct students to 12,089 and the number of distinct problems to 617, where students were randomized 62,638 times into treatment and 7,115 times into control.

## 6.3 Exploring effectiveness of Common Wrong Answer Feedback

Toward remediation of common wrong answers (CWAs) with the CWAFs, in the next step, we analyze the efficacy of CWAFs. We explore this by examining the binary correctness of the next problem. For our exploration of CWAFs, we use the pre-registered model [3] in our analysis plan to investigate the effectiveness of CWAFs. The pre-registered model is listed in 1.

$$\begin{aligned} &logit(nextproblemcorrectness \sim \\ &treatment * priorpercentcorrect + hintusage + \\ &attemptcount + (1|CWAwriter) + (1|problem) + (1|class)) \end{aligned} \quad (1)$$

We examine the effectiveness of CWAFs by interacting the treatment with average student performance on the previous 5 problems prior to working on the treatment problem, total hint usage, and attempts made on the treated problem. Additionally, we introduce the CWA writer, treatment problem unique identifier, and the class identifier of the student as a random intercept. The CWA writer is introduced as a random intercept to examine the variance in the effectiveness of the CWAFs across different teachers who wrote the feedback. The problem is introduced as a random intercept to control for the variance at a problem level that can be attributed to the problem's difficulty. Finally, the class is introduced as a random intercept as student motivation, and learning behavior are often influenced by their relative position with respect to their peers in the same class.

In this analysis, we investigate the effectiveness of CWAF in remediating gaps in student knowledge by using binary next-problem correctness as a dependent measure. CWAFs are designed to address and remedy the common misconceptions that the students exhibit. However, within the study, students are typically working on assignments with multiple problems that may or may not belong to the same set of Common Core Standards (or the same skill set). As these CWAFs address errors related to a specific set of skills, we hypothesize that the likelihood of knowledge transfer is higher for consecutive problems with the same set of skills. In contrast, the transfer of knowledge might be less likely for consecutive problems focusing on different sets of skills. As such, we perform two separate analyses examining the transfer of knowledge: 1) Between consecutive problems with the same set of common core standards and 2) Between consecutive problems in the same assignment irrespective of their common core standards. The results of these analyses are presented in the following section.

[3]The study has been pre-registered following open-science practices at BLINDED-URL

**Table 5: Exploring the effectiveness of CWAF by using next problem correctness(binary) as a dependent measure for the same set of Common Core Standards in consecutive problems.**

| | next_problem_correctness_binary | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.41 | 1.27 – 1.58 | **<0.001** |
| control treatment assignment [CWAF_treatment] | 1.07 | 1.01 – 1.14 | **0.015** |
| prior 5pr avg correctness | 6.14 | 5.12 – 7.38 | **<0.001** |
| attempt count | 0.91 | 0.90 – 0.92 | **<0.001** |
| hint count | 0.82 | 0.80 – 0.84 | **<0.001** |
| control treatment assignment [CWAF_treatment] × prior 5pr avg correctness | 0.79 | 0.66 – 0.96 | **0.017** |
| **Random Effects** | | | |
| $\sigma^2$ | 3.29 | | |
| $\tau_{00\ class\_xid}$ | 0.14 | | |
| $\tau_{00\ problem\_id}$ | 0.94 | | |
| $\tau_{00\ CWA\_writer}$ | 0.00 | | |
| ICC | 0.25 | | |
| $N_{problem\_id}$ | 617 | | |
| $N_{CWA\_writer}$ | 19 | | |
| $N_{class\_xid}$ | 1075 | | |
| Observations | 69209 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.074 / 0.303 | | |

*6.3.1 Between Consecutive Problems with the same set of Common Core Standards.* For the problems within the same set of common core standards within the consecutive problems, the results from the regression analysis are reported in table 5. We observe that students in the treatment condition are 7% more likely to answer the next problem correctly for the problems with the same set of common core standard tags. Additionally, students who make more attempts or ask for hints are less likely to benefit from the CWAF and answer the next problem correctly. While CWAFs do appear to have a net positive benefit, the model indicates that students with higher prior knowledge are less likely to answer the next problem correctly when exposed to CWAFs.

**Table 6: Exploring the effectiveness of CWAF by using next problem correctness(binary) as a dependent measure within the same assignment irrespective of the set of Common Core Standards associated with consecutive problems.**

| | next_problem_correctness_binary | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.35 | 1.24 – 1.47 | **<0.001** |
| control treatment assignment [CWAF_treatment] | 1.03 | 0.98 – 1.08 | 0.228 |
| prior 5pr avg correctness | 5.17 | 4.46 – 5.99 | **<0.001** |
| attempt count | 0.94 | 0.93 – 0.95 | **<0.001** |
| hint count | 0.86 | 0.84 – 0.87 | **<0.001** |
| control treatment assignment [CWAF_treatment] × prior 5pr avg correctness | 0.87 | 0.74 – 1.01 | 0.074 |
| **Random Effects** | | | |
| $\sigma^2$ | 3.29 | | |
| $\tau_{00\ class\_xid}$ | 0.12 | | |
| $\tau_{00\ problem\_id}$ | 0.85 | | |
| $\tau_{00\ CWA\_writer}$ | 0.00 | | |
| ICC | 0.23 | | |
| $N_{problem\_id}$ | 1018 | | |
| $N_{CWA\_writer}$ | 19 | | |
| $N_{class\_xid}$ | 1203 | | |
| Observations | 104747 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.061 / 0.276 | | |

*6.3.2 Between Consecutive Problems in the same Assignment irrespective of Common Core Standards.* For the problems irrespective of the common core standards within the consecutive problems, the results from the regression analysis are reported in table 6. We observed similar results on the other covariates; however, we did not observe a significant difference between students in control and treatment, indicating that the transfer of knowledge in consecutive problems with different sets of common core standards may or may not be observed.

## 7 DISCUSSION AND FUTURE WORKS

Our analysis found that incorrect answers on problems in ENY and IM repeat across school years as different groups of students from each school year made similar incorrect answers when working on the same problems. While the same CWAs were not the most common for the same problems in every school year, there was an obvious pattern indicating an overlap in the top 3 CWAs across school years. We also observed that teachers using IM and ENY prefer to assign "Practice Problems" over "Exit Tickets" for ENY. Similarly, teachers preferred to assign "Practice Problems" over "Student Facing Task" and "Cool Down" for IM as assignments. This claim was reinforced during the experimental analysis of CWAFs as we only generated CWAFs for CWAs of problems in "Practice Problems" of IM and ENY for grade 7. We observed that the CWAFs were randomized 623,857 times since their implementation in April '22. While various prior works exploring CWAs in the past have expressed concern regarding the reliability of CWAs as students in smaller samples often presented different CWAs on similar problems in studies exploring CWAs attributed to the "bugs" present in the students' synthesis of solutions [7, 40]. However, our analysis of CWAs substantiates the prevalence of CWAs. A potential cause of the replication challenges encountered by prior works [41] exploring the reliability of CWAs could be attributed to the smaller sample size at the prevalence of CWAs is pronounced and consistent at scale. It is important to note that our work does not claim to provide insight into the various underlying mechanisms students utilize when synthesizing solutions that can result in the incorrect answer due to "bugs" in their processes, but rather through this work, we aim to establish the reliability of the CWAs that can be caused by gaps in student knowledge, misconceptions, guess, slip, error, or bugs when formulating solutions.

While the primary objective of this paper was to explore the fidelity of CWAFs in this paper, we also wanted to focus on various design and development techniques that can potentially benefit future research. While the Learning@Scale community at large has designed and successfully developed systems at scale, it is somewhat concerning how we, as a community put little emphasis on documenting the various design and development principles that informed the successful implementation of such systems. As such, in this paper, we leverage the design philosophy commonly used in visualization projects to conduct goal and task abstraction that can elucidate the various aspects of the tool that are fundamental in the overall successful adoption of the tool. In our case, the objective was to develop a tool to augment teacher ability to examine CWAs when writing CWAFs. The primary benefit of the goals and task analysis is the ability to identify critical features a tool should facilitate and the hierarchy of these features to ensure the successful implementation of the tool. As such, in this paper, we present the fundamental goals and tasks a crowdsourcing tool needs to facilitate a successful adoption. Each goal is designed to build on prior goals and further enhance the process of facilitating crowdsourcing. While there is no evidence to suggest that the design philosophy used in the development of this crowdsourcing tool led to the creation of more effective feedback, we did observe that the CWAFs lead to positive learning outcomes across consecutive problems focusing on the same skill set. This positive outcome is particularly important in the domain of CWAFs research as there is mixed evidence regarding the fidelity of CWAFs, with some reporting positive results [26, 27, 41]. In contrast, others have reported on the lack of benefit in using CWAFs [35]. A well-designed system can provide powerful affordance that can enhance the quality of the outcome by facilitating exploration, learning, and collaboration when leveraging crowdsourcing.

In our final analysis, we examine the effectiveness of CWAFs by examining the transfer of knowledge on the next problem using the

binary correctness of the next problem as a dependent measure. We observe that the student benefit from CWAFs and are more likely to perform better on consecutive problems when working on the same set of skills. This finding in the context of IM and ENY curricula is particularly interesting; later problems focusing on the same set of skills within a problem set are generally designed to be more difficult than the prior problems. In comparison, the next problem performance within the assignment, irrespective of the set of skills associated with the problem, was not significant. However, rather interestingly, the effect of the CWAFs was not significantly different between the two regression models. Further analysis is required to develop our understanding of the usage of CWAFs to understand the underlying mechanism influencing knowledge transfer. While the focus of the problems within a problem set can differ, they are not entirely unrelated; future work needs to examine if the CWAFs were not effective because the focus of the problems was drastically different from each other or, conversely, if the CWAFs are facilitating shallow learning resulting in the students performing well on similar next problems without actually learning the concept addressed by the problem. We aim to build on our findings from this paper and further investigate student behavior around CWAFs. Additional work is required to gain insight into the productive usage of CWAFs. Prior works exploring student behavior around help [15] have explored the use of response time decomposition in inferring student effort in help usage.

Similarly, others have explored the correlation between structure, simplicity, and length of feedback and learning outcomes[Blinded for Review]. During our exploration of the effectiveness of the CWAFs, we observed that the variance in the model due to the CWA writer was negligible, indicating that the training and use of moderators to generate a consistent set of CWAFs that was based on the principle of Haitie et al. [] as demonstrated in **??**was successful. In future work, we intend to leverage the CWAFs generated through moderated crowdsourcing as a baseline when comparing the effectiveness of different CWAs. As these CWAFs were generated across 1,660 problems, this provides us with opportunities to test the effectiveness of different types of feedback across different topics and subfields of mathematics, e.g., geometry, statistics, algebra, and arithmetic.

While the focus of this paper is the exploration of CWA and the feasibility of crowdsourcing feedback from teachers to remediate the gaps in student knowledge that resulted in the CWAs through CWAF, we implore researchers and developers in our community of L@S to utilize our findings in the task abstraction that informed the design and development of our crowdsourcing tool and follow suit in documenting the design approaches they took in the developing their systems at scale to help inform future research.

## 8 CONCLUSION

At the onset of this research, we posited the validity of the idea of CWAs. In the subsequent sections, we presented evidence supporting that CWAs exist across problems and can be established at scale. We substantiate our claim by demonstrating how CWAs repeat across academic years despite shifts in the underlying student population. We leveraged this information to generate and collect CWAFs from teachers through the development of a crowdsourcing tool. Teachers using the crowdsourcing tool to generate CWAFs resulted in better learning outcomes, and there was evidence of knowledge transfer across consecutive problems focusing on the same set of skills. Further, we observed no significant effect of CWAFs when the consecutive problems focused on different sets of skills.

## 9 ACKNOWLEDGEMENT

Anonymized for Review

## REFERENCES


[1] Robert L Bangert-Drowns, Chen-Lin C Kulik, James A Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of educational research* 61, 2 (1991), 213–238.

[2] Robert L Bangert-Drowns, James A Kulik, and Chen-Lin C Kulik. 1991. Effects of frequent classroom testing. *The journal of educational research* 85, 2 (1991), 89–99.

[3] Sami Baral, Anthony F Botelho, John A Erickson, Priyanka Benachamardi, and Neil T Heffernan. 2021. Improving Automated Scoring of Student Open Responses in Mathematics. *International Educational Data Mining Society* (2021).

[4] Sameer Bhatnagar, Nathaniel Lasry, Michel Desmarais, and Elizabeth Charles. 2016. Dalite: Asynchronous peer instruction for moocs. In *Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13-16, 2016, Proceedings 11*. Springer, 505–508.

[5] Anthony F. Botelho, Sami Baral, John A. Erickson, Priyanka Benachamardi, and Neil T. Heffernan. (In Press) 2023. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning* ((In Press) 2023).

[6] Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2376–2385.

[7] John Seely Brown and Richard R Burton. 1978. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive science* 2, 2 (1978), 155–192.

[8] John Seely Brown and Kurt VanLehn. 1980. Repair theory: A generative theory of bugs in procedural skills. *Cognitive science* 4, 4 (1980), 379–426.

[9] Richard R Burton. 1982. Diagnosing bugs in a simple procedural skill. *Intellinget Tutoring Systems* (1982), 157–184.

[10] Jenny Yun-Chen Chan, Erin R Ottmar, and Ji-Eun Lee. 2022. Slow down to speed up: Longer pause time before solving problems relates to higher strategy efficiency. *Learning and Individual Differences* 93 (2022), 102109.

[11] Linda S Cox. 1975. Diagnosing and remediating systematic errors in addition and subtraction computations. *Arithmetic Teacher* 22, 2 (1975), 151–157.

[12] Ryan SJ d Baker, Albert T Corbett, Sujith M Gowda, Angela Z Wagner, Benjamin A MacLaren, Linda R Kauffman, Aaron P Mitchell, and Stephen Giguere. 2010. Contextual slip and prediction of student performance after use of an intelligent tutor. In *International conference on user modeling, adaptation, and personalization*. Springer, 52–63.

[13] Shayan Doroudi, Joseph Williams, Juho Kim, Thanaporn Patikorn, Korinn Ostrow, Douglas Selent, Neil T Heffernan, Thomas Hills, and Carolyn Rosé. 2018. Crowdsourcing and education: Towards a theory and praxis of learnersourcing. International Society of the Learning Sciences, Inc.[ISLS].

[14] Cornelia S Große and Alexander Renkl. 2007. Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and instruction* 17, 6 (2007), 612–634.

[15] Ashish Gurung, Anthony F Botelho, and Neil T Heffernan. 2021. Examining Student Effort on Help through Response Time Decomposition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 292–301.

[16] John Hattie. 1999. Influences on student learning. *Inaugural lecture given on August* 2, 1999 (1999), 21.

[17] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[18] 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.

[19] Thomas T Hills. 2015. Crowdsourcing content creation in the classroom. *Journal of Computing in Higher Education* 27, 1 (2015), 47–67.

[20] Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin* 119, 2 (1996), 254.

[21] Cheng-Fei Lai. 2012. Error Analysis in Mathematics. Technical Report# 1012. *Behavioral Research and Teaching* (2012).

[22] Nancyruth Leibold and Laura Marie Schwarz. 2015. The art of giving online feedback. *Journal of Effective Teaching* 15, 1 (2015), 34–46.
[23] Richard S Lysakowski and Herbert J Walberg. 1982. Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal* 19, 4 (1982), 559–572.
[24] Steven MOORE, Huy NGUYEN, and John STAMPER. 2020. Utilizing Crowdsourcing and Topic Modeling to Generate Knowledge Components for Math and Writing Problems. In *Proceedings of the 28th International Conference on Computers in Education*. 31–40.
[25] Tamara Munzner. 2009. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 921–928.
[26] Susanne Narciss. 2004. The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. *Experimental psychology* 51, 3 (2004), 214.
[27] Susanne Narciss. 2013. Designing and evaluating tutoring feedback strategies for digital learning. *Digital Education Review* 23 (2013), 7–26.
[28] Bobby Ojose. 2015. *Common misconceptions in mathematics: Strategies to correct them*. University Press of America.
[29] Bobby Ojose. 2015. Students' Misconceptions in Mathematics: Analysis of Remedies and What Research Says. *Ohio Journal of School Mathematics* 72 (2015).
[30] Thanaporn Patikorn. 2021. *Improvement on Hint and Explanation Crowdsourcing Method for an Online Learning Platform*. Ph. D. Dissertation. WORCESTER POLYTECHNIC INSTITUTE.
[31] Thanaporn Patikorn and Neil T Heffernan. 2020. Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 115–124.
[32] Ethan Prihar, Thanaporn Patikorn, Anthony Botelho, Adam Sales, and Neil Heffernan. 2021. Toward Personalizing Students' Education with Crowdsourced Tutoring. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 37–45.
[33] Amy Rummel and Richard Feinberg. 1988. Cognitive evaluation theory: A meta-analytic review of the literature. *Social Behavior and Personality: an international journal* 16, 2 (1988), 147–164.
[34] Sheryl J Rushton. 2018. Teaching and learning mathematics through error analysis. *Fields Mathematics Education Journal* 3, 1 (2018), 1–12.
[35] Lauren C Schnepper and Leah P McCoy. 2013. Analysis of misconceptions in high school mathematics. *Networks: An Online Journal for Teacher Research* 15, 1 (2013), 625–625.
[36] Douglas Selent and Neil Heffernan. 2014. Reducing student hint use by creating buggy messages from machine learned incorrect processes. In *International conference on intelligent tutoring systems*. Springer, 674–675.
[37] Raymund Sison and Masamichi Shimura. 1998. Student modeling and machine learning. *International Journal of Artificial Intelligence in Education (IJAIED)* 9 (1998), 128–158.
[38] Russell J Skiba, Ann Casey, and Bruce A Center. 1985. Nonaversive procedures in the treatment of classroom behavior problems. *The Journal of Special Education* 19, 4 (1985), 459–481.
[39] Gershon Tenenbaum and Ellen Goldring. 1989. A meta-analysis of the effect of enhanced instruction: Cues, participation, reinforcement and feedback and correctives on motor skill learning. *Journal of Research & Development in Education* (1989).
[40] Kurt VanLehn. 1982. Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. *The Journal of Mathematical Behavior* (1982).
[41] Kurt VanLehn, Stephanie Siler, Charles Murray, Takashi Yamauchi, and William B Baggett. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction* 21, 3 (2003), 209–249.
[42] Melanie R Weaver. 2006. Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education* 31, 3 (2006), 379–394.
[43] John Woodward and Lisa Howard. 1994. The misconceptions of youth: Errors and their mathematical meaning. *Exceptional Children* 61, 2 (1994), 126.
[44] Richard M Young and Tim O'Shea. 1981. Errors in children's subtraction. *Cognitive Science* 5, 2 (1981), 153–177.
[45] Mengxiao Zhu, Ou Lydia Liu, and Hee-Sun Lee. 2020. The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education* 143 (2020), 103668.