How Common are Common Wrong Answers? Crowdsourcing Remediation at Scale

Ashish Gurung
Worcester Polytechnic Institute
Worcester, Massachusetts, U.S.A.
agurung@wpi.edu

Sami Baral
Worcester Polytechnic Institute
Worcester, Massachusetts, U.S.A.
sbaral@wpi.edu

Morgan P. Lee
Worcester Polytechnic Institute
Worcester, Massachusetts, U.S.A.
mplee@wpi.edu

Adam C. Sales
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
asales@wpi.edu

Aaron Haim
Worcester Polytechnic Institute
Worcester, Massachusetts, U.S.A.
ahaim@wpi.edu

Kirk P. Vanacore
Worcester Polytechnic Institute
Worcester, Massachusetts, U.S.A.
kpvanacore@wpi.edu

Andrew A. McReynolds
Worcester Polytechnic Institute
Worcester, Massachusetts, U.S.A.
aamcreynolds@wpi.edu

Hilary Kreisberg
Lesley University
Cambridge, Massachusetts, U.S.A.
hkreisbe@lesley.edu

Cristina Heffernan
The ASSISTments Foundation
Auburn, Massachusetts, U.S.A
cristina.heffernan@assistments.org

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, Massachusetts, U.S.A.
nth@wpi.edu

ABSTRACT

Solving mathematical problems is cognitively complex, involving strategy formulation, solution development, and the application of learned concepts. However, gaps in students' knowledge or weakly grasped concepts can lead to errors. Teachers play a crucial role in predicting and addressing these difficulties, which directly influence learning outcomes. However, preemptively identifying misconceptions leading to errors can be challenging. This study leverages historical data to assist teachers in recognizing common errors and addressing gaps in knowledge through feedback. We present a longitudinal analysis of incorrect answers from the 2015-2020 academic years on two curricula, Illustrative Math and EngageNY, for grades 6, 7, and 8. We find consistent errors across 5 years despite varying student and teacher populations. Based on these Common Wrong Answers (CWAs), we designed a crowdsourcing platform for teachers to provide Common Wrong Answer Feedback (CWAF). This paper reports on an in vivo randomized study testing the effectiveness of CWAFs in two scenarios: next-problem-correctness within-skill and next-problem-correctness within-assignment, regardless of the skill. We find that receiving CWAF leads to a significant increase in correctness for consecutive problems within-skill. However, the effect was not significant for all consecutive problems within-assignment, irrespective of the associated skill. This paper investigates the potential of scalable approaches in identifying Common Wrong Answers (CWAs) and how the use of crowdsourced CWAFs can enhance student learning through remediation.

This work is licensed under a Creative Commons Attribution International 4.0 License.

L@S '23, July 20–22, 2023, Copenhagen, Denmark
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0025-5/23/07.
https://doi.org/10.1145/3573051.3593390

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → *Interactive systems and tools.*

KEYWORDS

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

ACM Reference Format:
Ashish Gurung, Sami Baral, Morgan P. Lee, Adam C. Sales, Aaron Haim, Kirk P. Vanacore, Andrew A. McReynolds, Hilary Kreisberg, Cristina Heffernan, and Neil T. Heffernan. 2023. How Common are Common Wrong Answers? Crowdsourcing Remediation at Scale. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3573051.3593390

1 INTRODUCTION

The intricacies of learning mathematics are cognitively complex. Solving math problems demands students to understand the problem's requirements and demonstrate their knowledge and comprehension of the topic [44]. Often, the problem-solving process involves breaking down the task into smaller sub-tasks that span several underlying concepts [7, 42]. This synthesis stage includes practicing various mathematical syntaxes, rules, and operations. The practice of synthesizing solutions reinforces students' knowledge and comprehension of the underlying concepts, thereby facilitating the development and consolidation of their understanding of mathematical principles [24, 47].

While the learning and synthesis processes may seem intuitive and straightforward, their analysis presents significant challenges [45]. The learner's individual problem-solving steps are intrinsic and can be challenging to deconstruct. Students can apply their inherent cognitive abilities to adopt different approaches towards solution synthesis [11, 46]. These approaches can vary, for example, in the complexity of the broken-down sub-task or the order in which the sub-tasks are solved [8].

Despite variations in approach, a fundamental understanding of mathematical processes is essential for problem-solving. However, gaps in knowledge, misconceptions, or "slips" can lead to incorrect responses [13]. Alternatively, insufficiently understood concepts may prompt students to guess answers or adopt incorrect problem-solving strategies, leading to a different set of errors [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 mathematical problems is critical for guiding the design of effective instructional practices to help students learn correct mathematical processes and problem-solving strategies [34].

The diagnosis and examination of "Common Wrong Answers" (CWAs) is critical to understand learning processes in the context of mathematics. CWAs can be used to enhance 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). For consistency, we will reference them as CBLP throughout this paper.

In a previous study, the authors of this paper examined the efficacy of two distinct types of Common Wrong Answer Feedback (CWAF)–verbose and detailed versus short and concise (c.f., [18]). The study employed a randomized control trial, where the control was business as usual, with no CWAF. The CWAs were proactively identified using a diagnostic model approach [7], and teachers, alongside learning activity designers, were tasked with generating the corresponding CWAFs. The analysis led to interesting insights for students working on mastery-based activities. The verbose and detailed feedback detailing both correct and incorrect steps undertaken by the students was detrimental to the student's likelihood of achieving mastery. On the other hand, short and concise CWAFs, while not significant, hinted towards a positive trend in facilitating student mastery.

In this current paper, we build on prior research by broadening our analysis of CWAs. We leverage historical data on a CBLP by analyzing 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 engineering a crowdsourcing platform that teachers can use to write 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. 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:

RQ 1 Do students commonly make similar errors when working on math problems?

RQ 2 What fundamental goals and tasks must a crowdsourcing platform provide when facilitating the generation of CWAF?

RQ 3 Does the remediation of CWAs with CWAFs lead to better learning outcomes?

2 BACKGROUND

2.1 Common Wrong Answers

Wrong answers are mistakes or errors that students typically make due to buggy rules, misconceptions about the topic, or gaps in knowledge. These CWAs have been the subject of substantial research in the fields of cognitive science and mathematical learning [7–9, 35, 57, 58].

Prior research [12, 43] has explored the correction of these common errors through instructional strategies. For instance, Brown et al., (1978) [7] analyzed frequent student errors when solving multi-digit subtraction problems and developed a diagnostic model that detects and elucidates these errors. Building on this, Brown et al., (1980) [8] introduced the "generative theory of bugs," a set of formal principles devised to explain the prevalent errors in procedural skills.

In their study, Sison et al., (1998) [48] proposed student modeling techniques to identify common errors in student work. They emphasized the need to assemble a "bug library," a collection of the most common misconceptions or errors made by a specific student population. However, they acknowledged the challenges in creating these libraries, as misconceptions vary depending on the student population, and different student groups may demonstrate unique types of misconceptions when solving mathematical problems.

In addition to the principles of learning theory and cognitive skill acquisition, research has also investigated the potential of algorithmically identifying common student misconceptions to rectify incorrect and buggy processes in students' work [36, 43]. Selent et al., (2014) [43] employed machine learning methods to predict CWAs and their underlying causes. They examined the effectiveness of providing *buggy messages* when a student makes a CWA. Their data suggested that these *buggy messages* led to a reduction in help-seeking behavior on a CBLP, indicating a possible rectification of common errors in students' work.

2.2 Feedback Intervention

Feedback is a significant factor influencing learning outcomes and achievement. However, the impact of feedback is contingent on its

[1] The data and code used in this paper are shared through open-science practices at https://github.com/AshishJumbo/LatS_CWAF

type and mode of delivery. Previous research on Feedback Interventions (FI) through meta-analyses has produced mixed results regarding their effectiveness on student performance [1, 2, 20, 26, 29, 39, 49, 50]. These results have spurred further research to explore the intricacies of FI, culminating in the development of Feedback Intervention Theory (FIT) [26]. FIT posits that FIs aim to capture the recipient's attention across three hierarchically organized levels: task learning, task motivation, and meta-task. While there are concerns about the general effectiveness of FIs [20], these concerns are less significant in an educational context as they have been found to be more beneficial in instructional settings. In a comprehensive synthesis of over 500 meta-analyses on the effects of schooling, Hattie (1999) (c.f., [20]) identified FIs as among the top 10 most influential factors on student achievement, thereby underscoring their effectiveness in promoting learning.

Effective feedback can help learners track their progress, validate their efforts, reinforce their progress, and impact their reactions and behavior when working on activities [10, 19, 59]. Feedback is indeed crucial to the student's learning experience, but the quality of the feedback varies greatly. The effectiveness of feedback is often influenced by student perception. Some studies have reported on constructive feedback from instructors to be the most beneficial [54]. Conversely, if the feedback was too vague or lacked content, its usefulness would diminish. Studies, such as [28], discuss how 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 for the remediation of common errors, CWAs, in students' work. We adopt the Hattie et al. (2007) [21] conceptualization of feedback[2], that expanded upon the generalized FIT model and proposed a theoretical model aiming 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. [21] for enhancing learning. The 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 questions by operating across four levels of instruction: (a) task level, (b) process level, (c) self-regulation level, and (d) self-level. Therefore, effective feedback should recognize if the task requirement is understood, demonstrate the correct processes required to complete the task, include instructions that direct the learner towards the next productive actions, and include evaluation and affect (usually positive) to personalize the instruction.

2.3 Common Wrong Answer Feedback

Prior research has dedicated significant focus to the remediation of common errors in students' work [32, 33]. A study by Vanlehn et al. (2003) [52], for instance, evaluated the interplay between expert human tutors and physics students, specifically examining the efficacy of tutor explanations in rectifying student errors. The study reported that only certain explanations led to improved learning, with the effectiveness of feedback heavily contingent on the content and the question at hand. Moreover, shorter and more precise explanations were observed to be more effective than their longer, more elaborate counterparts. Thus reinforcing our prior work exploring CWAFs, where long and verbose CWAFs were detrimental to student mastery rates on mastery-based activities [18].

Additional studies have indicated the limitations of guided instructions in rectifying errors originating from misconceptions of previously learned skills [41]. These findings suggest that deeply ingrained misconceptions and errors might pose substantial difficulties to rectify over time.

Further research has proposed the use of error analysis methods as an essential step towards understanding students' ability to identify and explain errors in problems [17, 27, 40]. These studies involved presenting students with erroneous examples and requiring them to identify and articulate the errors within them. In particular, Rushton et al. (2018) [40] reported that this approach to error analysis led to better knowledge retention compared to traditional methods of learning mathematics.

2.4 Crowdsourcing Instruction

Crowdsourcing has emerged as a prevalent method in K-12 education for gathering feedback on instructional materials [16, 25, 55]. Leveraging various authoring tools, educators can create and disseminate educational content that is more representative. A variety of CBLPs and tools have integrated the crowdsourcing approach to encourage instruction and teacher-authored content [4, 15, 22, 37, 53, 56].

Research underscores the potential of crowdsourcing in enriching online learning experiences. It enables on-demand teacher support, tutoring, provision of hints, and explanations [14, 23, 30, 37, 38, 56]. Moreover, several studies have explored the use of crowdsourcing to collect teacher-given scores and feedback messages (instructive guidance) for students' answers on open-ended math problems to develop automated grading and feedback generation using Natural Language Processing (NLP) algorithms [3, 5]. The effectiveness of crowdsourcing in enhancing instructional materials and student learning experiences on online platforms has been well-documented [37, 38].

Building on these insights, our current study aims to crowdsource CWAFs by developing a platform for teachers to identify and rectify CWAs.

3 EXPLORING COMMON WRONG ANSWERS

To answer **RQ 1**, we explored the commonality of CWAs by examining data from students in grades 6, 7, and 8 who worked on problems in two commonly used curricula for mathematics in the US: Illustrative Mathematics (IM) and EngageNY (ENY) over a five-year period from '15-'16 to '19-'20. The students' data were collected from ASSISTments [22] learning platform. 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

[2] *[21] 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 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.*

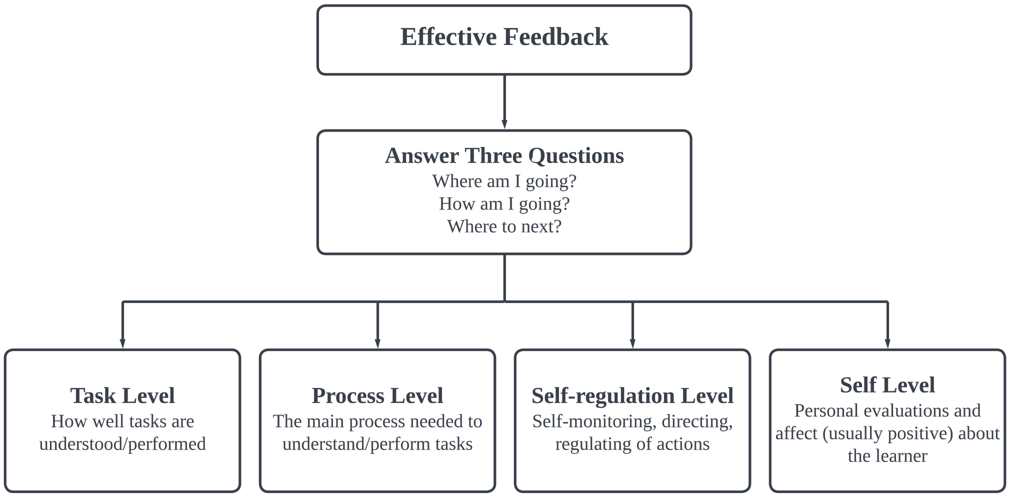

Effective Feedback
Answer Three Questions
Where am I going?
How am I going?
Where to next?
Task Level
How well tasks are understood/performed
Process Level
The main process needed to understand/perform tasks
Self-regulation Level
Self-monitoring, directing, regulating of actions
Self Level
Personal evaluations and affect (usually positive) about the learner

Figure 1: A model of feedback for enhanced learning, taken from Hattie et al. (2007) [21]

more than 20 students in at least one of the 5 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.

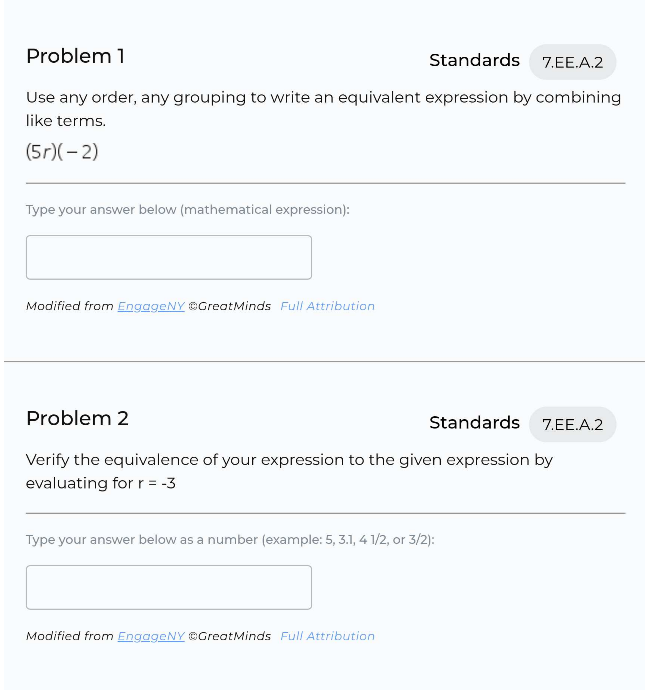
In the ASSISTments platform, students are typically assigned a sequence of problems, each of which may or may not involve the same set of skills as defined by the Common Core Standards.

Figure 2 provides an example from the EngageNY (ENY) curriculum, where two consecutive problems are associated with the same Common Core Standards, hence demanding a similar skill set. The first problem calls for the simplification of an equation, while the second entails verifying the results derived from the initial problem. Problems sharing a common skill set, like the ones mentioned, offer a greater likelihood of knowledge transfer compared to those derived from different Common Core Standards.

In our investigation of incorrect response frequency, we analyzed each student's initial incorrect attempt on problems, facilitating the generation of the top three CWAs for each problem. To enhance the reliability of the CWAs, we added an additional criterion: where we only considered the problems that had been attempted by at least 20 students during the school year, with more than 10 students producing the most common incorrect answer.

In our analysis, we found that 1,045 problems had CWAs spanning at least two academic years. Table 2 provides an example of these CWAs across academic years for the second problem presented in figure 2, from ENY grade 7 module 3 lesson 1. As reported in table 2, we observe that the first CWA met the commonality threshold in four out of the 5 academic years, indicating consistency. However, the second and third CWAs demonstrated some fluctuation, with ranks interchanging in some years, and entirely new CWAs appearing in others.

Additionally, we noticed a declining trend in the number of students across the school years. This decline can be attributed to a version upgrade to the CBLP used in our analysis. During the '18-'19 academic year, teachers began transitioning to the newer version. Although this change reduced the total number of students available for our analysis in the later academic years, it did not hinder our ability to demonstrate the prevalence of CWAs. The same CWAs reappeared despite changes in the student and teacher populations working on the problems.


Problem 1 Standards 7.EE.A.2

Use any order, any grouping to write an equivalent expression by combining like terms.

$(5r)(-2)$

Type your answer below (mathematical expression):

Modified from EngageNY ©GreatMinds Full Attribution

Problem 2 Standards 7.EE.A.2

Verify the equivalence of your expression to the given expression by evaluating for r = -3

Type your answer below as a number (example: 5, 3.1, 4 1/2, or 3/2):

Modified from EngageNY ©GreatMinds Full Attribution

Figure 2: An example of two consecutive problems from ENY Grade 7 Module 3 Lesson 1 where both problems have the same set of Common Core Standards.

Our exploratory analysis of the occurrence of CWAs revealed a pattern of repetition across academic years. A more in-depth analysis of the problems featuring CWAs indicated that the majority

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 two or more academic years.

	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 on the second problem as presented in figure 2. The threshold for the CWA requirement was met in 4 of the 5 academic years from '15-'20. The threshold required more than 20 students to work on the problem in each academic year with more than 10 students making the same CWA.

School Year	Number of Students	Incorrect Count	Correct Answer	First CWA		Second CWA		Third CWA	
				Answer	Count	Answer	Count	Answer	Count
'15 - '16	214	62	30	-30	42	5	5	13	2
'16 - '17	354	75	30	-30	44	-17	3	-13	5
'17 - '18	332	98	30	-30	71	-17	5	0	3
'19 - '20	243	63	30	-30	38	-15	4	-17	4

of the problems belonged to *"Practice Problems"* (in IM) and *"Problem Sets"* (in ENY)[3]. As the term problem set is generally used to represent a set of problems that can be assigned to students, we will refer to both *Practice Problems* and *Problem Sets* activities as *Practice Problems* throughout this paper.

In the following section, we detail an iterative process of goal and task analysis. This process guided the design and development of a crowdsourcing tool intended for teachers. The tool's aim is to facilitate the creation of CWAFs that can address and remediate the gaps in students' understanding that resulted in the CWAs.

4 TASK ABSTRACTION

Toward answering **RQ 2**, in this section we detail our process for designing and developing a crowdsourcing tool, which involved consulting with experienced teachers, teacher trainers, domain experts, and researchers exploring similar tools. Our analysis comprises two main parts: a goals analysis, which involved creating a hierarchy of goals that the tool should facilitate, and a task analysis, which focused on defining low-level tasks.

During the goals analysis, we broke down each goal into a series of sub-goals that directly align with teacher needs. For instance, a high-level goal might be: facilitate effective feedback, which could be broken down into sub-goals such as 'analyze student error rates' and 'allow teachers to easily input their feedback'. We utilize the sub-goals to identify the visualization components needed in the crowdsourcing tool to meet teachers' needs effectively. We utilized the "Nested Model for Visualization" (c.f., [31]), a common Human-Computer Interaction (HCI) technique, to identify the fundamental goals of a crowdsourcing tool.

Upon validating the high-level goals and sub-goals with end-users and domain experts, we proceeded with task analysis, defining low-level tasks allowing browsing, exploring, and identifying various aspects of the data to facilitate the sub-goals. These tasks, derived from the Brehmer and Munzner topology (c.f., [6]), provided a useful roadmap for designers and developers during the tool's creation. While our crowdsourcing tool doesn't include the elaborate visualization components associated with common HCI projects, the Nested Model for Visualization, and Brehmer and Munzner's topology proved invaluable in identifying the tool's fundamental goals and tasks, which ultimately helped in enhancing teachers' ability to formulate effective feedback.

After conducting several iterations of goal and task analyses for further refinement of the goals and tasks, we present the final version of the goals and tasks used to develop our tool in the following sub-sections.

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 processes the students might have taken 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 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

[3] IM and ENY have different types of activities in their curricula. IM has 3 types of activities *"Practice Problems"*, *"Student Facing Tasks"* and *"Cool Down"* and ENY has 2 types of activities *"Problem Sets"* and *"Exit Tickets"*

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 throughout the task abstraction processes emphasized the importance of goal 3 in fostering self-actualization for teachers through collaborative feedback enhancement. It enriches their participation in a generation of CWAFs through peer support and fostering a sense of camaraderie. Such opportunities allows teachers to contribute to and benefit from the collective knowledge.

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 [6].

Table 4 illustrates high-level tasks that can guide the design and development of features in the crowdsourcing tool, facilitating one or more sub-goals. Together, these tasks contribute to achieving the main goals of the crowdsourcing project. While these tasks can be further decomposed into more specific sub-tasks, we focus only on high-level tasks to avoid unnecessary complexity. We believe these tasks are self-explanatory and refrain from extensive elaboration to conserve space and prevent redundancy.

It's worth noting that this list is not exhaustive; it's a reference derived from our interaction with teachers and other stakeholders during the tool's design and development phase. It provides insights into what we found useful but should not be considered as an all-encompassing guide to creating an effective crowdsourcing tool. In fact, it is our hope that future work in the field of crowdsourcing makes amendments or modifications to this list based on their unique project requirements and insights.

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 prior section. In order to facilitate the fundamental goals described in table 3 we designed a new crowdsourcing platform within the ASSISTments ecosystem. 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 displays the teacher perspective on a problem set in IM curricula for grade 7, unit 8, lesson 8–based on the common core standard for "Probability and Sampling". As the figure illustrates a teacher has analyzed the first CWA for the problem and provided appropriate CWAF. The teacher can substantiate the CWAs, **Goal 1**, 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.

Beyond examining the validity of the CWAs the teacher can also explore other problems in the problem set and their CWAs to gain insight into how students have historically struggled within the problem set. The ability to explore previous and consecutive problems in the problem set can contextualize the CWAF more effectively, facilitating **Goal 2**. We posit that such insights substantiating and contextualizing the CWAs, coupled with peer collaboration and review, **Goal 3**, will enhance the generation of effective CWAFs.

The primary focus of this paper is to analyze CWAs and evaluate the efficacy of CWAFs in addressing the underlying causes of the CWAs. We collaborated with 24 experienced middle school teachers using IM or ENY in their classrooms. These teachers were tasked with generating CWAFs for Grade 7 *Practice Problems.* To ensure the feedback aligned with the curriculum requirements, teachers received preliminary training from domain experts. The experts also offered continuous feedback and served as moderators during the crowdsourcing process to maintain the quality of CWAFs. After the CWAFs were crowdsourced, the experts performed a final review to approve the feedback, marking it as ready for student use.

In the following section, we detail a randomized control trial conducted at the student problem level to evaluate the efficacy of CWAFs at scale.

6 IMPLEMENTING COMMON WRONG ANSWER FEEDBACK

The crowdsourced CWAFs, once approved by the moderators, were integrated into ASSISTments. The initial implementation, which took place in April '22, has since evolved through various iterations. As of now, crowdsourced CWAFs for 1,660 problems are provided to students working on problems whenever they make a CWA.

6.1 Experimental Design

Once the students start a problem, students are randomized into either a control group, business-as-usual (no CWAF), or a treatment group (receiving CWAFs). Ideally, randomizing students once they make a CWA would be 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 beforehand and analyze the effectiveness of CWAFs on the treated group. We implemented a 90:10 randomization split, providing a 90% chance of a student being assigned to treatment and a 10% chance to control. This ratio was strategically chosen to optimize access to learning opportunities for as many students as possible.

6.2 Dataset

Since the initial implementation of the first batch in April '22, CWAFs have been randomized across 20,044 students working on

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.

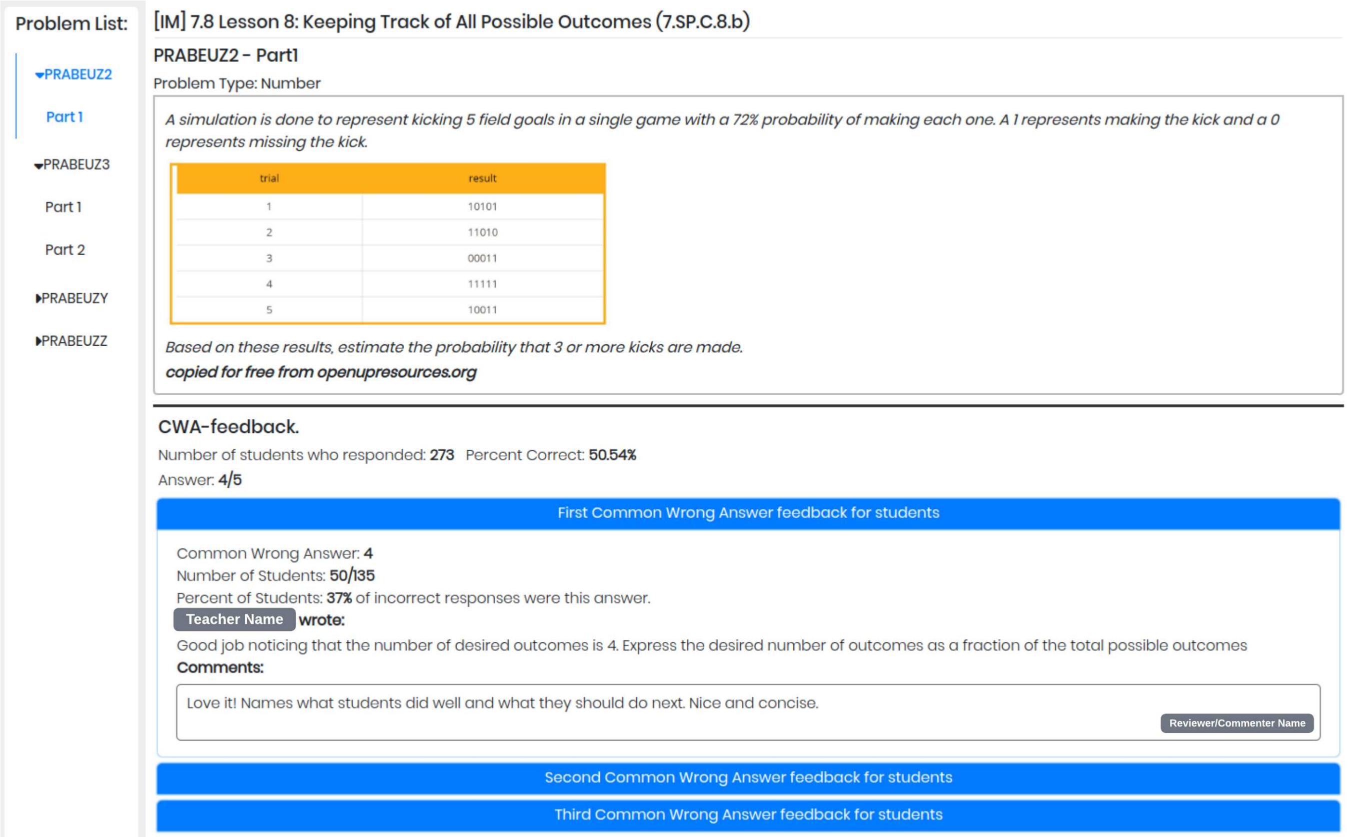
Table 4: Task analysis deconstructing the feature requirements of each sub-goal.

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 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.

1,387 problems in ENY and IM a total of 623,857 times; students were assigned 560,897 times to treatment and 62,960 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 CWA 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 14,672 unique students who were randomized and made at least one CWA when working across 947 problems. With this, we have 96,398 instances of students randomized to treatment and 10,960 to control. As we used a 90:10 randomization design, we explored the balance across conditions by conducting a binomial hypothesis test on the next problem attempt after receiving a CWAF. Our sample failed the binomial hypothesis test indicating an imbalance across the attrition rates for treatment and control, as such we scored 0s for instances where the students dropped out without attempting the next problem. While this data is for students working on problems within the same problem set, different problems within a single problem set can have different sets of common core standards. 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 12,175 and the number of distinct problems to 535, where students were randomized 62,688 times into treatment and 7,080 times into control.

6.3 Evaluating the Effectiveness of Common Wrong Answer Feedback

For answering **RQ 3**, in this section we analyze the efficacy of CWAFs in the remediation of common wrong answers (CWAs). We explore this by examining the binary correctness of the next problem using the *lme4* package in R. We use a pre-registered



[IM] 7.8 Lesson 8: Keeping Track of All Possible Outcomes (7.SP.C.8.b)

PRABEUZ2 - Part1

Problem Type: Number

A simulation is done to represent kicking 5 field goals in a single game with a 72% probability of making each one. A 1 represents making the kick and a 0 represents missing the kick.

trial	result
1	10101
2	11010
3	00011
4	11111
5	10011

Based on these results, estimate the probability that 3 or more kicks are made.

copied for free from openupresources.org

CWA-feedback.

Number of students who responded: **273** Percent Correct: **50.54%**

Answer: **4/5**

First Common Wrong Answer feedback for students

Common Wrong Answer: **4**

Number of Students: **50/135**

Percent of Students: **37%** of incorrect responses were this answer.

Teacher Name **wrote:**

Good job noticing that the number of desired outcomes is 4. Express the desired number of outcomes as a fraction of the total possible outcomes

Comments:

Love it! Names what students did well and what they should do next. Nice and concise.

Reviewer/Commenter Name

Second Common Wrong Answer feedback for students

Third Common Wrong Answer feedback for students

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.

logistic regression model to explore the effectiveness of CWAFs [4]. The pre-registered logistic regression model is listed in equation 1.

$$\begin{aligned} next\ problem\ correctness \sim treatment\ * \\ prior\ 5\ problem\ avg\ correctness \\ +\ (1|CWA\ writer)\ +\ (1|problem)\ +\ (1|class) \end{aligned} \tag{1}$$

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. Rather than employing the more commonly used average prior percent correct, this study uses the average correctness of the last 5 problems. As the running average can be more sensitive to fluctuation in students' performance, likely attributable to the error rates that can occur when learning a new concept. Using a running average enables the model to effectively capture instances where the student is optimally positioned to benefit form receiving a CWAF.

In addition, we introduce the identifiers for the CWA writer, the specific problem being treated, and the student's class as random intercepts in our model. The CWA writer is included to examine potential variations in the effectiveness of CWAFs across different teachers who provided the feedback. The specific problem identifier is included to control for variance at the problem level that may be attributable to various problem related factors including difficulty, guess- and slip-rates. Finally, the class identifier is used to account for the impact of classroom-level factors, as students' motivation and learning behaviors are often influenced by their relative standing among their classmates.

The analysis aims to explore our initial hypothesis that knowledge transfer is more likely for consecutive problems focusing on the same set of skills. Therefore, we conduct two separate analyses: 1) Between consecutive problems with the same set of common core standards (within-skill) and 2) Between consecutive problems in the same assignment (within-assignment), regardless of their common core standards.

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 (within-skill), the results from the regression analysis are reported in table 5. We observe that students in the treatment condition had significantly higher odds to answer the next problem correctly for the problems with the same set of common core standard tags (Odds-Ratio = 1.07, p-value = 0.028). The fixed effect of mean-centered prior 5 problem average correctness was significant and highly predictive

[4] The study has been pre-registered following open-science practices at https://osf.io/wp2a7

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 (within-skill) in consecutive problems.

	next problem correctness binary		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.93	0.83 – 1.04	0.189
CWAF treatment	1.07	1.01 – 1.13	**0.028**
prior 5 problem avg correctness	6.25	5.21 – 7.49	**<0.001**
CWAF treatment × prior 5 problem avg correctness	0.80	0.66 – 0.97	**0.021**
Random Effects			
σ^2	3.29		
$\tau_{00\ \text{class xid}}$	0.15		
$\tau_{00\ \text{problem id}}$	1.02		
$\tau_{00\ \text{CWA writer}}$	0.00		
$N_{\text{problem id}}$	533		
$N_{\text{CWA writer}}$	19		
$N_{\text{class xid}}$	1072		
Observations	69632		
Marginal R^2 / Conditional R^2	0.073 / NA		

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

	next problem correctness binary		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.92	0.85 – 1.00	0.052
CWAF treatment	1.03	0.99 – 1.08	0.166
prior 5 problem avg correctness	5.14	4.44 – 5.95	**<0.001**
CWAF treatment × prior 5 problem avg correctness	0.88	0.76 – 1.03	0.102
Random Effects			
σ^2	3.29		
$\tau_{00\ \text{class xid}}$	0.14		
$\tau_{00\ \text{problem id}}$	0.87		
$\tau_{00\ \text{CWA writer}}$	0.00		
$N_{\text{problem id}}$	943		
$N_{\text{CWA writer}}$	19		
$N_{\text{class xid}}$	1201		
Observations	107084		
Marginal R^2 / Conditional R^2	0.048 / 0.272		

of next-problem-correctness. While CWAFs do appear to have a net positive benefit, there was a significant interaction between treatment and prior 5 problem average correctness indicating a potential heterogeneous treatment effect[5].

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 (within-assignment), the results from the regression analysis are reported in table 6. We observed similar results on the other covariates; however, while leaning in the positive direction we did not observe a significant difference between students in control and treatment, indicating that the transfer of knowledge in consecutive problems to be inconclusive (Odds-Ratio = 1.03, p-value = 0.188). The fixed effect of mean-centered prior 5 problem average correctness was significant and highly predictive of next-problem-correctness, however the interaction between treatment and prior 5 problem average correctness while similar to the within-skill analysis was not significant[6].

7 DISCUSSION AND FUTURE WORKS

Our analysis revealed a relative consistency in the incorrect answers made by students across academic years. 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. We also observed that teachers using IM and ENY prefer to assign *Practice Problems* over *Exit Tickets*, *Student Facing Task*, and *Cool Down* problem sets. While various prior works exploring CWAs in the past have expressed concerns regarding the reliability of CWAs [7, 51], our analysis substantiates the commonality of CWAs. A potential cause of the replication challenges encountered by prior works [52] exploring the reliability of CWAs could be attributed to the smaller sample size, as our analysis does indicate the prevalence of CWAs 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, 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 be potentially beneficial to future research. While the Learning@Scale (L@S) community at large has designed and successfully developed systems at scale, it is noteworthy that there has been a limited emphasis within our community on documenting the various design and development principles that inform the successful implementation of such systems. As such, in this paper, we leverage the design philosophy commonly used in visualization projects to conduct

[5] There were 2 problems in the within-skill dataset that only had students in treatment and none in control which resulted in the problem ids being dropped

[6] There were 3 problems in the entire treated dataset that only had students in treatment and none in control which resulted in the problem ids being dropped

task abstraction that can elucidate the various aspects of crowdsourcing that are fundamental in the overall successful adoption of such tools. In our case, the objective was to develop a tool that can augment teacher ability to examine CWAs when writing CWAFs. The primary benefit of the goals and task analysis is to identify critical features a tool should facilitate and the hierarchy of such features to ensure the successful implementation of the tool. As such, this paper presents 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 in comparison to other design philosophies, 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 [32, 33, 52]. In contrast, others have reported on the lack of benefit in using CWAFs [18, 41]. 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.

As attested by the lack of variance in the outcome due to CWA writer, as random intercepts, in both the within-skill and within-assignment models reported in table 5 and table 6 respectively. This observation suggests that the training and use of moderators to generate a consistent set of CWAFs, following the principles outlined by Hattie et al. (2007) [21] as presented in figure 1, was successful. In future work, we intend to leverage the CWAFs generated through moderated crowdsourcing as a baseline when comparing the effectiveness of different CWAF designs. As these CWAFs were generated across 1,660 problems, we can now hypothesize and test the effectiveness of different types of feedback across different topics and subfields of mathematics, e.g., geometry, statistics, algebra, and arithmetic.

In our final analysis, we examine the effectiveness of CWAFs by examining the transfer of knowledge on the next problem using the binary measure of the next-problem-correctness in two contexts, within-skill, and within-assignment. Our findings reveal that students appear to benefit from CWAFs, as evidenced by their increased likelihood of solving consecutive problems correctly within-skill. This outcome is noteworthy, particularly in the context of IM and ENY curricula, where subsequent problems within a skill set tend to increase in difficulty. However, we did not observe a similar benefit on subsequent problems within-assignment. These findings suggest a contextual aspect of the effectiveness of CWAFs. Further investigation is needed to develop our understanding of these dynamics. For instance, while the within-skill knowledge transfer could occur due to the CWAFs effectively addressing student needs, it is also entirely plausible that the CWAFs are causing shallow learning–as evidenced by the lack of knowledge transfer within-assignments. Additionally, further analysis exploring learner behavior around CWAFs is required to understand if students are attentive to the CWAFs. A prior analysis has explored student attention towards hints by utilizing response time decomposition, where higher attention to hints was correlated with student learning outcomes [19].

While the focus of this paper has been the exploration of CWA and the efficacy of crowdsourced feedback, we implore fellow researchers and developers in our L@S community to consider leveraging similar task abstraction methodologies in their own work. We believe the insights provided in our goal analysis, presented in Table 3, can serve as initial guardrails for informing future research aimed at developing tools exploring similar crowdsourcing challenges. Such methodologies can potentially streamline the process of identifying the fundamental features in crowdsourcing contexts, thus enhancing overall efficiency and output.

8 CONCLUSION

At the onset of this research, we posited the existence and prevalence of CWAs in a learning context. Our findings substantiate our initial hypothesis, revealing a remarkable persistence of CWAs across different academic years, even with changing student populations. Utilizing this understanding, we successfully developed a new crowdsourcing tool to facilitate the collection of Common Wrong Answer Feedbacks (CWAFs) from educators. Our analysis demonstrates that the integration of these teacher-generated CWAFs leads to improved learning outcomes, particularly evidenced by the observed transfer of knowledge across consecutive problems that focus on the same skill set (within-skill). Interestingly, the effectiveness of CWAFs was less pronounced when consecutive problems irrespective of associated skill sets (within-assignment). This distinction offers a promising avenue for further investigation in future studies. Furthermore, our work has produced a baseline that can be leveraged by future research exploring CWAFs.

9 ACKNOWLEDGEMENT

We would like to thank NSF (*e.g.*, 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (*e.g.*, R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (*e.g.*, P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), and Schmidt Futures. None of the opinions expressed here are that of the funders.

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. 2023. Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning* (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] Jere Confrey. 1990. Chapter 8: What constructivism implies for teaching. *Journal for Research in Mathematics Education. Monograph* 4 (1990), 107–210.

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

[13] 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.

[14] 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.

[15] Paul Denny, Andrew Luxton-Reilly, and John Hamer. 2008. The PeerWise system of student contributed assessment questions. In *Proceedings of the tenth conference on Australasian computing education-Volume 78.* Citeseer, 69–74.

[16] 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].

[17] 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.

[18] Ashish Gurung, Sami Baral, Kirk P Vanacore, Andrew A Mcreynolds, Hilary Kreisberg, Anthony F Botelho, Stacy T Shaw, and Neil T Hefferna. 2023. Identification, Exploration, and Remediation: Can Teachers Predict Common Wrong Answers?. In *LAK23: 13th International Learning Analytics and Knowledge Conference.* 399–410.

[19] 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.

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

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

[22] 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.

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

[24] Carolyn Kieran. 1981. Concepts associated with the equality symbol. *Educational studies in Mathematics* 12 (1981), 317–326.

[25] Juho Kim et al. 2015. *Learnersourcing: improving learning with collective learner activity.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[26] 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.

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

[28] Nancyruth Leibold and Laura Marie Schwarz. 2015. The art of giving online feedback. *Journal of Effective Teaching* 15, 1 (2015), 34–46.

[29] 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.

[30] 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.

[31] Tamara Munzner. 2009. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 921–928.

[32] 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.

[33] Susanne Narciss. 2013. Designing and evaluating tutoring feedback strategies for digital learning. *Digital Education Review* 23 (2013), 7–26.

[34] Bobby Ojose. 2015. *Common misconceptions in mathematics: Strategies to correct them.* University Press of America.

[35] Bobby Ojose. 2015. Students' Misconceptions in Mathematics: Analysis of Remedies and What Research Says. *Ohio Journal of School Mathematics* 72 (2015).

[36] Zachary Pardos, Scott Farrar, John Kolb, Gao Xian Peh, and Jong Ha Lee. 2018. Distributed representation of misconceptions. International Society of the Learning Sciences, Inc.[ISLS].

[37] 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.

[38] 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.

[39] 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.

[40] Sheryl J Rushton. 2018. Teaching and learning mathematics through error analysis. *Fields Mathematics Education Journal* 3, 1 (2018), 1–12.

[41] 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.

[42] Alan H Schoenfeld. 2016. Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (Reprint). *Journal of education* 196, 2 (2016), 1–38.

[43] 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.

[44] Robert Siegler. 2009. Implications of cognitive science research for mathematics education. *Colección Digital Eudoxus* 8 (2009).

[45] Robert S Siegler. 1984. Strategy choices in addition and subtraction: How do children know what to do? *Origins of cognitive skills* (1984).

[46] Robert S Siegler. 1988. Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child development* (1988), 833–851.

[47] Robert S Siegler. 1998. *Emerging minds: The process of change in children's thinking.* Oxford University Press.

[48] Raymund Sison and Masamichi Shimura. 1998. Student modeling and machine learning. *International Journal of Artificial Intelligence in Education (IJAIED)* 9 (1998), 128–158.

[49] 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.

[50] 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).

[51] Kurt VanLehn. 1982. Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. *The Journal of Mathematical Behavior* (1982).

[52] 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.

[53] Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. UpGrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale.* 1–10.

[54] Melanie R Weaver. 2006. Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education* 31, 3 (2006), 379–394.

[55] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing.* 405–416.

[56] 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.* 379–388.

[57] John Woodward and Lisa Howard. 1994. The misconceptions of youth: Errors and their mathematical meaning. *Exceptional Children* 61, 2 (1994), 126.

[58] Richard M Young and Tim O'Shea. 1981. Errors in children's subtraction. *Cognitive Science* 5, 2 (1981), 153–177.

[59] 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.