Identification, Exploration, and Remediation: Can Teachers Predict Common Wrong Answers?

Ashish Gurung
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
agurung@wpi.edu

Sami Baral
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
sbaral@wpi.edu

Kirk P. Vanacore
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
kpvanacore@wpi.edu

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

Hilary Kreisberg
Lesley University
Cambridge, Massachusetts, USA
hkreisbe@lesley.edu

Anthony F. Botelho
University of Florida
Gainesville, Florida, USA
abotelho@coe.ufl.edu

Stacy T. Shaw
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
sshaw@wpi.edu

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

ABSTRACT

Prior work analyzing tutoring sessions provided evidence that highly effective tutors, through their interaction with students and their experience, can perceptively recognize incorrect processes or "bugs" when students incorrectly answer problems. Researchers have studied these tutoring interactions examining instructional approaches to address incorrect processes and observed that the format of the feedback can influence learning outcomes. In this work, we recognize the incorrect answers caused by these buggy processes as Common Wrong Answers (CWAs). We examine the ability of teachers and instructional designers to identify CWAs proactively. As teachers and instructional designers deeply understand the common approaches and mistakes students make when solving mathematical problems, we examine the feasibility of proactively identifying CWAs and generating Common Wrong Answer Feedback (CWAFs) as a formative feedback intervention for addressing student learning needs. As such, we analyze CWAFs in three sets of analyses. We first report on the accuracy of the CWAs predicted by the teachers and instructional designers on the problems across two activities. We then measure the effectiveness of the CWAFs using an intent-to-treat analysis. Finally, we explore the existence of personalization effects of the CWAFs for the students working on the two mathematics activities.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**; **E-learning**.

KEYWORDS

Causal Inference, Buggy Messages, Common Wrong Answers, Automated Feedback

ACM Reference Format:
Ashish Gurung, Sami Baral, Kirk P. Vanacore, Andrew A. McReynolds, Hilary Kreisberg, Anthony F. Botelho, Stacy T. Shaw, and Neil T. Heffernan. 2023. Identification, Exploration, and Remediation: Can Teachers Predict Common Wrong Answers?. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3576050.3576109

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03.
https://doi.org/10.1145/3576050.3576109

1 INTRODUCTION

Learning mathematics is a cognitively complicated process. For many mathematics-based questions designed to help students practice math syntax, rules, and operations, students may demonstrate their knowledge by applying procedural skills to synthesize solutions. Analyzing the synthesis processes can be particularly challenging as the underlying mechanisms of the individual steps taken to reach a solution are not obvious. As a result of gaps in student knowledge or misconceptions, students may make errors on one or more steps in solving a problem due to a misconception or "slip" [8] that can lead to a variety of potential incorrect answers. Conversely, gaps in student knowledge or shallowly-learned concepts may cause students to guess at answers or otherwise apply the wrong approach, resulting in an entirely different set of incorrect answers. Regardless the cause, the experience of errors during problem-solving without directed feedback as to how to rectify those errors may impede a student's learning progress. Understanding the common errors that are experienced by students as they interact with math problems is critical for guiding the design of effective instructional practices to help students learn correct mathematical processes and problem-solving strategies. The diagnosis and examination of "Common Wrong Answers" (CWAs) is important for understanding learning processes in the context of mathematics, and may be utilized to develop better educational technologies that, in conjunction with teachers, can better meet

the needs of individual students–educational technologies often referenced as Computer Aided Learning Platform (CALP), Online Learning Platform (OLP), or Intelligent Tutoring Systems (ITS).

Despite the complexity of the synthesis process in mathematics learning, teachers' knowledge of mathematics and ability to anticipate areas of potential difficulty or struggle among their students is correlated with student learning outcomes [13]. Within this, many teachers are able to use experiential knowledge to recognize the types of mistakes, sometimes referred to as "bugs" [3], and misconceptions produced by their students. Researchers have explored teacher approaches modeling student knowledge states by deconstructing their process and reverse engineering such models for procedural skills in mathematics (c.f., [3]). Brown and colleagues ([3]) investigated the use of procedural networks in constructing diagnostic models. These models provided teachers and instructional designers with learning and assessment value. A deeper understanding of the incorrect processes causing the incorrect answers can be leveraged in designing a more effective learning and assessment activity [18]. A fundamental takeaway from the diagnostic model is the recognition of many teachers' ability to address faulty processes performed by the students that result in these incorrect responses. However, not all these bugs and faulty processes can be addressed and adequately explained by teachers. The task of diagnosing the students' errors in itself is a procedural skill that is challenging and is often susceptible to misidentification by the teachers [3, 25]. Furthermore, describing these common processes can be complicated as several different incorrect processes can generate the same outcome resulting in misjudgment when justifying and addressing such student misconceptions. Therefore, proper tools and methods are essential to facilitate the diagnosis and analysis of CWAs. With the analysis and diagnosis of these CWAs, it is equally important to address the cause of these CWAs effectively. We can address student needs through tailored instructions to avoid misconceptions or provide feedback/hints to the students as they make these common mistakes.

In this paper, we examine two experiments that were designed to leverage teachers' and instructional designers' ability to construct diagnostic models to identify common bugs in student processes, while working on problems, that resulted in CWAs. The teachers were also asked to construct Common Wrong Answer Feedback (CWAF) messages based on the inferred bugs in the diagnostic model that resulted in the CWAs. First, we explore the fidelity of proactively identifying CWAs by leveraging the diagnostic models. If the diagnostic models can help teachers and instructional designers correctly identify the majority of the CWAs, then a similar approach can be adopted by various educational technologies in the identification of CWAs and their remediation through CWAFs. Second, we measure the effectiveness of these CWAFs by examining the learning outcomes of students working on mastery-based assignments. We compare the mastery rates between students who receive a CWAF when making a CWA with those who don't receive CWAF. We posit that the use of CWAFs will enhance the student learning experience by helping them identify the bugs or address their misconceptions resulting in higher mastery rates. Finally, We extend our analysis to explore heterogeneous treatment effects to explore potential opportunities for personalized interventions for high- and low-performing students. While the primary objective of this work is to examine the efficacy of CWAFs in general, we additionally explore the benefits of two different design approaches by comparing the effectiveness of short and concise CWAFs against more elaborate CWAFs.

With this, the main research questions we address in the paper are:

RQ 1 Can teachers and instructional designers identify common wrong answers on math problems?
RQ 2 Does receiving common wrong answer feedback improve short-term learning outcomes?
RQ 3 Do high- and low-performing students benefit differently from common wrong answer feedback?

2 RELATED WORKS

In most mathematics-based questions, CWAs typically arise from a buggy rule, a lack of knowledge among the students, or a common misconception about the topic. There are various prior works investigating the common errors made by students during their mathematical thinking process [3–5, 19, 27, 28]. Others have also focused on rectifying these errors through instruction [7, 23]. As such, Brown and colleagues [3] analyzed students' incorrect responses to multi-digit subtraction problems to build a diagnostic model that helps detect and explain the incorrect responses in students' work. Furthermore, in [4], they explain the known/common bugs with a set of formal principles called the "generative theory of bugs," that transforms a procedural skill to generate all the possible buggy processes for that skill. Sison and colleagues [24] present several studies involving student modeling and to explain the significance of recognizing a "bug library" in student modeling tasks; this library is defined as the collection of the most common misconceptions or errors made by a population of students in the same domain. Further, they present the challenges in the construction of these libraries, as a different population of students may exhibit different types of bugs during the synthesis of mathematics solutions.

While the fundamental mechanism behind the CWAs is explained by the principles of learning theory and cognitive skill acquisition, various researchers have explored the likelihood of algorithmically identifying these buggy procedures to rectify the incorrect processes or buggy processes resulting in incorrect responses. A study from Selent et al. [23] proposes the use of machine learning techniques to predict CWAs and their causes in students' work and suggests using buggy messages to remedy these wrong answers. They further measured the reduction of help-seeking behavior (i.e. characterizing student learning as needing less help over time by the learning system) by leveraging these buggy messages within an online learning platform.

Various other researchers have explored the effectiveness of feedback in rectifying student errors [16, 17]. A study from Vanlehn and colleagues [26] observes the interaction between expert human tutors and physics students to study the effect of tutor explanations to address errors. This study found only some tutor explanations to be associated with improved learning when students exhibited difficulty, indicating that the effectiveness of the feedback varied with the content and the question. Furthermore, short and concise explanations were observed to be more effective in comparison to more elaborate explanations. Other research has identified an

inability of guided instructions to remediate errors emerging from student misconceptions among previously learned skills [22]; this suggests that deeply ingrained misconceptions may be more difficult to rectify over time. Other works [10, 15, 21] explored the use of error analysis methods by studying students' ability to identify and explain exhibited errors. These studies have explored presenting erroneous examples to students by asking them to detect and explain the error in the examples. Rushton et al. [21], report on the approach of error analysis leading to better knowledge retention over the traditional methods of learning mathematics.

3 METHODOLOGY

For all of our analyses, we utilize data that was collected from a randomized controlled trial designed to measure the learning impacts of CWAF. In this section, we first describe the study design and characteristics of the dataset. We then discuss the analysis conducted to address our first research question examining how well teachers and instructional designers can proactively identify the CWAs for two mathematics concepts. As teachers and instructional designers were asked to write CWAFs we then describe and report on the results of the randomized controlled trial to measure the impacts of CWAF on two short-term measures of learning. Finally, we examine interaction effects within this study to measure heterogeneous treatment effects among high- and low-performing students.

3.1 Study Design

Exploring the effectiveness of CWAFs uses two activities on the ASSISTmetns platform[12]; both problem sets have a mastery-based design which provides students with practice problems until they are able to demonstrate sufficient knowledge of the given concept. While some systems utilize a model-based measure of mastery using Knowledge Tracing [6] or similar approaches, the designers of the two activities in our analysis used an arbitrary threshold of N-Consecutive Correct Responses (N-CCR) with N = 3; that is, students must answer three consecutive problems correctly without the use of system-provided on-demand tutoring (e.g. hints), in order to complete the assignment. Kelly et al. [14] compared the performance of N-CCR (N=3) against a BKT model and found the performance of the two approaches to be comparable. Furthermore, Prihar et al. [20] have reported on studies extending the N-CCR experiments by exploring the benefits of N = 2, 3, 4, and 5 as thresholds and found N = 3 to be the optimal threshold for mastery-based math activities.

The instructional designers designed the content used in the study to align to the Common Core State Standards [1] for grade 7. The first activity focuses on the "Number System" (7.NS.A.3), and the second focuses on "Expressions & Equations" (7.EE.B.4). Students working on the activities get randomly assigned to a treatment or control condition–students in the treatment condition get feedback if their attempt is a CWA whereas the students in the control condition do not get any feedback. The students are assigned 10 random problems from a pool of ∼50 problems. Students in both conditions must answer 3 consecutive problems correctly to demonstrate mastery over the material. There is a daily limit of 10 problems per condition unless the student answers the 9^{th} or 10^{th} problem correctly; in such cases, the daily limit is extended to 11 and 12 problems, respectively. If the student cannot demonstrate mastery within the ten problems, they must wait until the next day to work on the problem set (this feature is intended to encourage students to seek help rather than continue to struggle on the assignment). Demonstrating mastery is the primary measure of success in both the activities, but also observe reaching this daily limit as a measure of wheel-spinning [2].

Instructional designers and teachers collaborated to design two problem templates per activity for both "2-Step Equations" and "Order of Operations" with the aim of generating problems that adequately addressed the objectives of the activities. Teachers can build problem templates in ASSISTments such that teachers can generate multiple problems using the same template. The templates used in generating the problems and an example per template are presented in 1. Teachers and instructional designers analyzed the generated problems to construct diagnostic models that postulate the approaches students could take when solving the problems along with the steps where bugs can occur in their approach due to "guess", "slip", or "misconception". The bugs were used to predict CWAs and generate templates for CWAFs. In the interest of preserving space and adhering to the conference's page limit, the templates for the CWAF and examples have been provided with the supplementary materials of this paper[1]. While we do not elaborate on the templates used in generating the CWAFs within this paper, we will briefly describe the two design approaches for CWAFs. As exemplified in figure 2, the students in the treatment condition of "2-Step Equations" activity get a CWAF when their attempt is a CWA, whereas students in the control condition do not get feedback. The CWAF consists of three main sections: (a) in blue, the core idea required to answer the problem; (b) in green, the correct steps the students likely took to synthesize an answer; and (c) in red, the crucial buggy step where the student made an error. Alternatively as shown in figure 3, the students in the treatment condition of "Order of Operations" activity get a CWAF that is more short and succinct in design. In our analysis of these two studies we explore the general effectiveness of CWAFs by analyzing their general effectiveness as well as exploring their effectiveness on their own as they have different designs. We analyze the two designs separately as prior works analyzing human tutor feedback in physics have suggested that a simpler and shorter explanations are more beneficial to students in contrast to more elaborate explanations resulting in the motto, "Ask more and tell less"[26].

3.2 Description of Dataset

The data was collected across 9 academic years and their respective summer sessions in the United States (the academic year 2013-14 to the Summer of 2022)[2]. During this period, the teachers accessed the two mastery-based activities as assignments for their students. Both activities fit the lesson plan as they align with Illustrative Math curricula under the Common Core Standards [1]. During this period, 587 middle school teachers in the United States assigned one or both mastery-based activities to 1283 of their classes resulting in

[1]The templates for the CWAFs are publicly available at *https://osf.io/gjst9/*
[2]The dataset and all the code used in this work is publicly available at *https://github.com/AshishJumbo/LAK_CWAF*

Order of Operations	2-Step Equations
Template 1: What is the solution to the expression below? %v{a} + %v{b} x %v{c} Example: What is the solution to the expression below? 7 + 4 x 3	Template 1: Solve for %v{a} %v{c1}%v{a} + %v{c2} = %v{c3} Example: Solve for a. 9a + 10 = 28
Template 2: What is the solution to the expression below? %v{a} - %v{b} x %v{c} Example: What is the solution to the expression below? 5 - 2 x 5	Template 2: Solve for %v{a} $\frac{\%v\{a\}}{\%v\{b\}} + \%v\{c\} = \%v\{d\}$ Example: Solve for y. $\frac{y}{2} + 6 = 4$

Figure 1: The two templates used to generate the problems across the two activities "Order of Operations" and "2-Step Equations" respectively along with an example for each template.

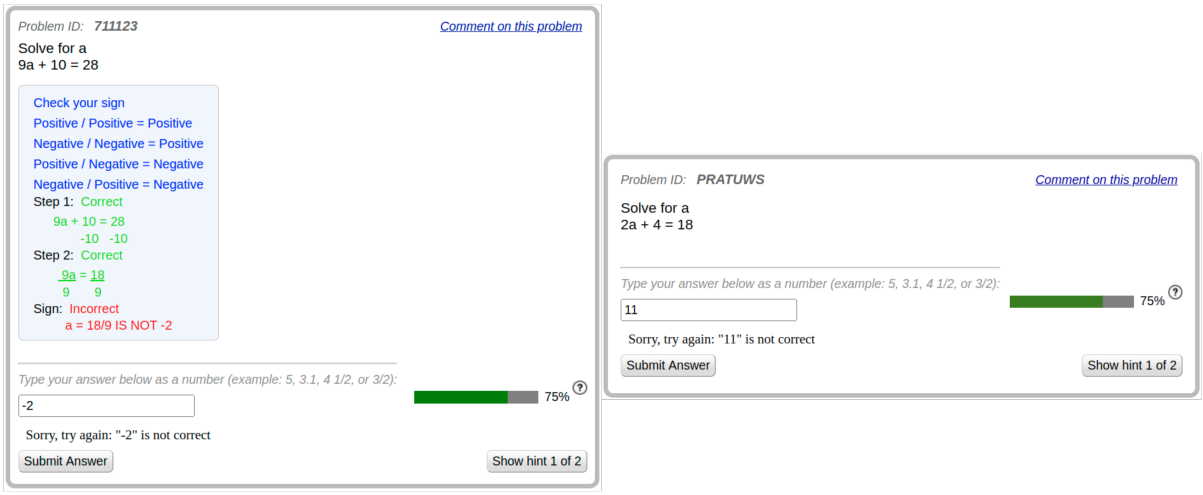

Figure 2: Example problems in treatment (problem on the left) and control (problem on the right) condition for "2-Step Equations" activity. The CWAF is provided to students when they provide a CWA in the treatment condition.

23,655 students working on the activity. The assignment-to-class ratio in the dataset is not one-to-one. Some teachers using the CALP prefer to divide their students into subgroups and assign them separate assignments within a single classroom. Another reason for the discrepancy in the one-to-one relationship is the Learning Tool Interoperability (LTI) integration within Canvas, a Learning Management System (LMS). School districts using Canvas occasionally group all students at a grade level into a single group and divide them into subgroups according to their classes. This grouping structure is problematic as the entire grade level now appears as a single class during LTI integration; this is a known issue with Canvas LTI integration.

As this is an in-vivo study, there were a few occasions where a teacher gave out the same activity to their student if their students initially performed poorly on the assignment. For instances where students worked on the mastery-based activity more than once,

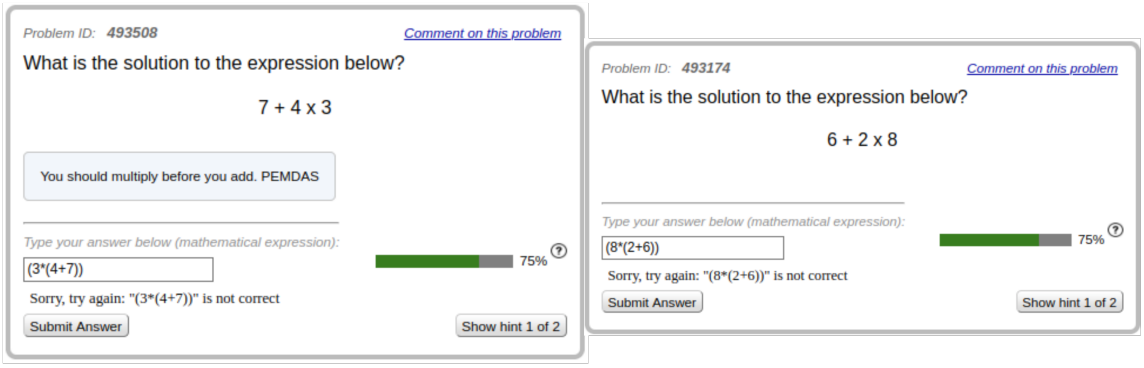

Figure 3: Example problems in treatment (problem on the left) and control (problem on the right) condition for "Order of Operations" activity. The CWAF is provided to students when they provide a CWA in the treatment condition.

we only analyzed the instances where the students worked on the activity for the first time and dropped all the other instances. Additionally, there were some instances where the students worked on both activities (*i.e.*, the students was assigned to the treatment in the first activity and control in the second activity)–in such scenarios, we dropped the student record for the second activity to avoid spillover effects within the study. Table 1 lists the number of teachers, classes, assignments, and students after implementing the filtration procedures on the data.

4 DEFINING COMMON WRONG ANSWERS

In this section, we analyze the incorrect answers provided by students while working on the mastery-based assignment and explore how common these common wrong answers actually are. We then extend our analysis to explore the ability of teachers and instructional designers in CALPs to predict CWAs. We analyzed all the instances when a student provided an incorrect response on their first attempt to help explore our first research question (RQ1) to evaluate teachers' and instructional designers' ability to anticipate and identify CWAs effectively. We limit our analysis to the first attempt, as all other attempts combine a corrective step to account for the incorrectness of the first attempt in the formulation of a solution.

Teachers and instructional designers analyzed mathematical problems using the Common Core State Standards and inferred diagnostic models of students synthesizing solutions to these problems. Table 2 presents the number of CWAs teachers and designers proactively predicted by analyzing the possible incorrect answers. The teachers and instructional designers analyzed the incorrect answers and processes for their likelihood of occurring based on experience and understanding of student approach to solving the problems. The incorrect answers that were considered the most likely were labeled CWAs. The teachers and instructional designers provided CWAFs to address the incorrect process that led to the CWAs. An example of a CWA and the associated CWAF is shown in the example provided in treatment problem in figures 2 & figure 3.

4.1 Identifying & Analyzing CWAs

The mastery-based activity had similar problems between treatment and control conditions, albeit not the same. In order to identify the CWAs, we analyzed all the first attempts where the students' answers were incorrect. As the aim is to explore the ability of instructional designers and teachers to leverage their teaching experience and insight into predicting the CWAs, we only analyze the problems in the treatment condition as the teachers had only predicted the CWAs for the treatment problems. We analyzed the CWAs using two arbitrary thresholds of N= 5 and 10–the answer is a CWA if N or more students submitted the answer.

Table 3 analyzes the CWAs across mastery-based activities where 5 or more students provide the incorrect answer. The instructional designers were able to predict CWAs for the problems in the "Order of Operations" where ~85% of the CWAs were correctly predicted and had associated feedback. Of the incorrect responses of the students, 2528 responses were CWAs with feedback from the instructional designers; however, only 2361 of the incorrect responses crossed the threshold of 5, indicating that certain incorrect messages were misclassified as common. Additionally, there were 81 instances where students provided CWAs were not identified by the teacher. Predicting CWAs for problems in the "2-Step Equations" was more challenging as only ~54% of the CWAs were correctly identified. Furthermore, identifying CWAs in "2-Step Equations" was more challenging as the teachers failed to identify 192 CWAs that occurred more than five times, resulting in 2037 instances where we failed to provide CWAFs.

Table 4 analyzes the CWAs across mastery-based activities using a higher threshold of 10 or more incorrect attempts. With a higher threshold, the instructional designers were more effective at predicting the CWAs for the problems in the "Order of Operations" activity. While the teachers accurately predicted all of the CWAs that occurred, the teachers identified 270 CWAs, of which only 57 (~21%) were common, *i.e.*, $N \geq 10$. Identifying CWAs for the problems in the "2-Step Equations" even at a higher threshold still presented challenges as only 143 (~72%) CWAs that occurred were

Table 1: Filtered list of teachers, classes, assignments and student working on the two problem set.

	Order of Operations	2-Step Equations	Combined
Teachers	202	458	587
Classes	386	954	1282
Assignments	497	954	1282
Students	6679	16976	23655

Table 2: CWAs identified by teachers by analyzing the problems.

	problems	Teacher Identified CWAs
Order of Operations	54	270
2-Step Equations	52	359

Table 3: Analyzing CWAs that were made by the students with a threshold N $\geq$ 5.

	Teacher Identified CWAs	Observed CWAs
Order of Operations		
CWAs identified by teacher	270	88
CWAs not identified by teacher	–	15
2-Step Equations		
CWAs identified by teacher	359	228
CWAs not identified by teacher	–	192

correctly identified. In contrast, teachers were unable to identify 54 (~21%) CWAs and provide appropriate CWAFs.

4.2 Results of Identifying CWAs

From our analysis of the CWAs using the arbitrary threshold of N = 5 or 10, we observed that the ability to predict CWAs varies across topics. While the instructional designers were more effective at predicting the CWAs for the "Order of Operations" compared to the "2-Step Equations", the general accuracy of the predicted CWAs was relatively low. When the threshold for commonality was 5: ~32% of the teacher predicted CWAs were actually made by the students working on the "Order of Operations", and teachers were unable to predict 15 of the new CWAs from students. For "2-Step Equations", ~63% of the teacher predicted CWAs were actually made by the students, and 192 new CWAs were observed which was not previously predicted by the teacher.

Likewise, when the threshold for commonality was 10: ~21% of the teacher predicted CWAs were made by the student for "Order of Operations," and students did not make any new CWAs on this problem set. For "2-Step Equations", ~39% of the teacher predicted CWAs were actually made by the students and 54 new CWAs were observed. While the instructional designers had some success in proactively identifying CWAs, upon accounting for the time and effort required to identify the CWAs and their inaccuracy, the approach taken in identifying CWAs in the paper appears to be highly inefficient. Further analysis and re-evaluation of the CWAs is required before exploring the utilization of CWAFs in math-based activities.

5 ANALYSIS OF THE EFFECTIVENESS OF CWAFS

In this section, we evaluate the effect of CWAFs relative to no CWAFs in helping students learn the underlying concept addressed in the problem sets to explore our second research question (RQ2). We hypothesize that the CWAFs will positively impact learning by helping students understand gaps in their knowledge. Our hypothesis is based on the intuition that students who make a CWA are closer to the answer. An appropriately designed CWAF has a higher likelihood of helping the student answer the problem, *i.e.*, recognizing the bug and reevaluating their answer formulation process can help the student answer the problem and learn from their mistakes. We examine student mastery and wheel-spinning learning outcomes for our analysis. Wheel-spinning is described as an unproductive learning behavior characterized by high student persistence while making very little progress towards mastering the given skill on concept [2]; analogous to a car getting stuck in the ice or mud, the student is "spinning their wheels" and applying effort to learn, but unable to make progress due to a gap in their knowledge. For our analysis, wheel-spinning is operationalized as students failing to exhibit mastery by answering 3 consecutive problems correctly before reaching the daily threshold of 10 problems.

5.1 Descriptive Statistics

We evaluated the student data on the mastery-based activities and compared the problems to mastery, hint usage, average problem

Table 4: Analyzing CWAs that were made by the students with a threshold N ≥ 10.

	Teacher Identified CWAs	Observed CWAs
Order of Operations		
CWAs identified by teacher	270	57
CWAs not identified by teacher	–	0
2-Step Equations		
CWAs identified by teacher	359	143
CWAs not identified by teacher	–	54

difficulty, and average student scores on the problems. This exploration was done to develop our intuition regarding the effect of CWAFs on mastery rates, average hint usage, problem difficulty, and average student performance on the assignment. Table 5 presents the descriptive statistics across conditions for the two activities. We observed that students in the treatment condition (CWAFs) of the "2-Step Equations", on average, needed more problems to reach mastery, asked for more hints, found the problems more difficult, and performed poorly. Simultaneously we observed that students in the treatment condition (CWAFs) of the "Order of Operations", on average, needed relatively more problems to reach mastery, asked for fewer hints, earned higher scores per problem, and had better performance. As the treatment and control problems were generated using a template, the problems are similar in structure. However, while the problems are similar, they could be different in difficulty; we cannot separate the effect of the CWAFs, the problem difficulty, or a combination of the two on students' performance on the assignment. From our exploration, we intuit that the two different designs of the CWAFs appear to have differing effects on student performance, with lower performance on the treatment condition of the "2-Step Equation" and higher performance on the treatment condition of the "Order of Operations". The CWAFs provided for "2-Step Equation" were more verbose, whereas the CWAFs provided for "Order of Operations" were short and concise.

5.2 Methods to Examine Effects of CWAF on Learning

To evaluate the effects of CWAF on student learning behaviors, we estimated mastery of knowledge component and wheel-spinning as learning outcomes using a series of multi-level logistic regressions. For each outcome, we ran three models, one of which included data from both the activities (2-step Equations and Order of Operations), and then two others analyzed the effect of CWAFs for each activity individually. This approach allows us to estimate the effect of CWAFs in general and separately for each activity as the two CWAFs have different designs, i.e., "2-Step Equations" had more elaborate CWAFs, whereas "Order of Operations" had moreconcise CWAFs. We included random intercepts for students' teachers as much of the variance in outcomes was associated with students' teachers. Prior to accounting for the treatment effects, the teacher accounted for the following variances in the learning outcomes: mastery (ICC = 0.37) and wheel-spinning (ICC = 0.27). The p-values of our analysis were adjusted using Benjamini-Hochberg to adjust for the potential inflation of false discovery rates due to multiple comparisons [9].

The logitsic regressions were estimated because mastery and wheel spinning are binary outcomes. Equation 1 is the base model used to address this research question. For any given assignments completed by student i, the equation for the likelihood of the outcome (mastery or wheel spinning) is 1 where γ_{00} is the fixed intercept and μ_{0t} is the random intercept for each teacher. CWAF_i is a binary indicator for whether a student is in the CWAF condition, and the coefficient for the effect of the CWAF problem sets condition is γ_{10}.

$$logit(\text{Outcome is True for Student } i \text{ with teacher } t) = \gamma_{00} + \gamma_{10}\text{CWAF}_i + \mu_{0t} \quad (1)$$

5.3 Results on the Effectiveness of CWAFs

Overall, we observed that the CWAFs significantly impacted both the likelihood that they exhibit mastery and the likelihood that they would wheel-spin. Figure 4 presents the treatment effects for each activity and learning outcome. In table 6 & table 7 we present our analysis exploring the effect of CWAFs on mastery and wheel-spinning behavior. CWAFs had an overall negative effect on the likelihood that students would master the knowledge component (γ_1 = -1.30, SE = 0.06, p = 0.027) and a positive effect on the likelihood that students would wheel-spin during the activity (γ_1 = 0.51, SE = 0.09, $p < 0.001$). Although the effects were significant for both outcomes when both activities were combined, the patterns of significance varied by activity. For the "2-Step Equations" activity, the effects of CWAF were for both mastery (γ_1 = -0.51, SE = 0.09, p = 0.001) and wheel-spinning (γ_1 = 0.21, SE = 0.06, $p < 0.001$). Yet, for the "Order of Operations" activity, neither of the effects on mastery (γ_1 = 0.22, SE = 0.14, p = 0.144) nor wheel-spinning (γ_1 = 0.30, SE = 0.27, p = 0.264) were significant. Notably, the point estimate for the CWAF effect on the likelihood of mastery was positive, along with most of the confidence interval. This suggests that a more precise estimate to form a future study may produce a positive result.

6 EXPLORING PERSONALIZATION EFFECTS

6.1 Identifying Heterogeneous Treatment Effects

To determine whether the effect of CWAF on mastery and wheel spinning differs based on students' general knowledge of math concepts, we added an interaction between students' prior percent

Table 5: Descriptive Statistics of the experiment across the control and treatment condition for the two activites.

	2-Step Equation				Order of Operations			
	Control		Treatment		Control		Treatment	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Average problems to mastery	4.28	2.66	4.68	3.52	4.48	1.95	4.53	2.11
Average total hints access on assignment	0.27	0.69	0.31	0.73	0.15	0.55	0.08	0.39
Average score per problem	0.82	0.08	0.79	0.06	0.82	0.10	0.85	0.08
Average student score (%)	87.67	20.43	86.04	20.27	86.39	21.12	88.38	19.45

Table 6: Effect of Common Wrong Answer Feedback (CWAF) on Mastery by Activity

	Both Activities		2-Step Equations		Order of Operations	
Predictors	*Log-Odds*	*SE*	*Log-Odds*	*SE*	*Log-Odds*	*SE*
Intercept	2.95***	0.08	2.68***	0.09	3.66***	0.17
CWAFs (Treatment)	-1.30**	0.06	-0.21**	0.06	0.22	0.14
Random Effects						
σ^2	3.29		3.29		3.29	
τ_{00}	1.93_t		1.83_t		1.68_t	
ICC	0.37		0.38		0.36	
N	587_t		458_c		202_c	
	23604_i		16926_i		6678_i	

Table 7: Effect of Common Wrong Answer Feedback (CWAF) on Wheel-Spinning by Activity

	Both Activities		2-Step Equations		Order of Operations	
Predictors	*Log-Odds*	*SE*	*Log-Odds*	*SE*	*Log-Odds*	*SE*
Intercept	-4.32***	0.10	-3.99***	0.11	-5.25***	0.29
CWAFs (Treatment)	0.51***	0.09	0.55***	0.09	0.31	0.28
Random Effects						
σ^2	3.29		3.29		3.29	
τ_{00}	1.18_t		1.03_t		0.76_t	
ICC	0.26		0.24		0.19	
N	587_t		458_c		202_c	
	23604_i		16926_i		6678_i	

correct in the CALP platform and the CWAF condition to the base model used in Section 5 (Equation 1). For students who completed problems in the CALP platform prior to working on the experiment, we have data on their prior performance i.e their prior percent correctness. We use students' average scores in these problems as an estimate of students' math ability. Prior percent correct was added as a standardized score to the model to improve interoperability. The standardization was calculated using group mean centering based on the activity (using the mean and standard deviation of the sample form each activity) as students in the activities had significantly different prior accuracy (t = 7.65, DF = 10941, p < 0.001).

Of the original sample, 21,793 students had completed at least ten (10) problems in the CALP before the experiment. We excluded students who had completed fewer than ten problems in the CALP platform prior to our study fewer than this amount of data would provide poor estimates of math ability. The exclusion criterion was balanced as 8.45% students from the CWAF condition and 6.98% students from the control condition were dropped. Therefore, the exclusion does not bias our estimates of the CWAF. The prior percent correct of this analytic sample ranged from 0% to 100% with a mean of 72.16% and a deviation of 14.07%.

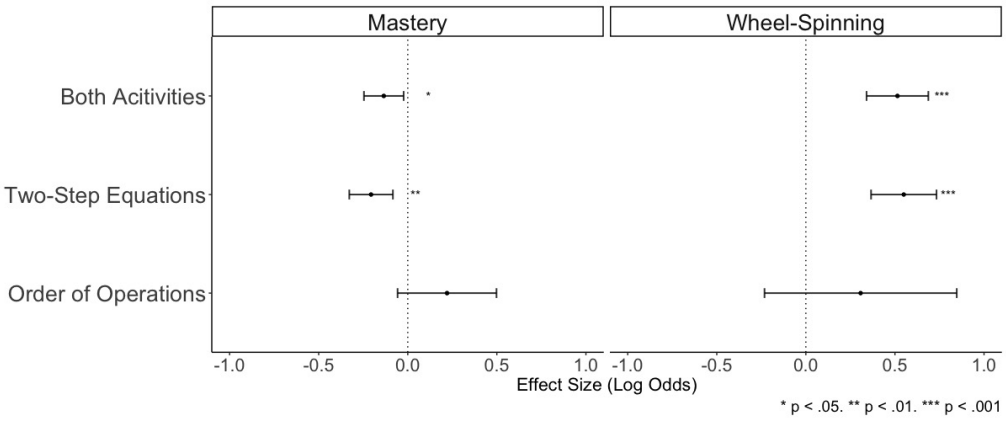

Figure 4: Comparing the effect of Common Wrong Answer Feedback (CWAFs) on mastery and wheel spinning behavior of students.

6.2 Results Exploring Personalization

Overall there was a significant interaction between students' prior percent correct and the CWAFs condition. Table 8 displays the results for these models. For students with the mean prior accuracy, the effect of CWAF was negative (γ_1 = -0.17, SE = 0.07, p = 0.017). The interaction effect was also negative (γ_2 = -0.11, SE = 0.05, p = 0.047), showing that the effect of CWAF was greater in the negative direction as students prior percent correct compared is higher.

When both the activities ("2-Step Equations" and "Order of Operations") were modeled separately, an interesting pattern emerged. For the "2-Step Equations" activity, the treatment effect was significant (γ_1 = -0.18, SE = 0.08, p = 0.019), but the interaction was not significant (γ_2 = -0.08, SE = 0.06, p = 0.192), showing that the CWAF had a consistently negative effect regardless of students prior percent correctness. Alternatively, for the Order of Operations activity, the main effect was not significant (γ_1 = -0.06, SE = 0.18, p = 0.756), but the interaction was significant (γ_2 = -0.36, SE = 0.14, p = 0.013). Figure 5 displays this interaction. Hence, in the "Order of Operations" activity, there was no significant effect of CWAF for students with average prior percent correct, but the treatment effect became greater in the negative direction for students with higher prior percent correct.

There were no significant interactions between the treatment effects and prior percent correct in any of the models predicting wheel-spinning. This is not surprising as the prevalence of wheel spinning is fairly low (described in detail in Section 5), and wheel spinning is more common among low-performing students with lower prior percent correct. Therefore it makes sense that the effect would not vary by prior percent correct.

7 DISCUSSION AND FUTURE WORKS

Our analysis did find that a substantial number of students commonly provide the same incorrect answer to problems. However, teachers can be inaccurate in identifying the CWAs and, from the randomized trial, the CWAFs did not seem to help address gaps in students' knowledge, on average. From our exploration of our first research question, we posit that further analysis is required in defining CWAs. The approach to proactively identifying CWAs seems inefficient and inaccurate, even for experienced mathematics teachers and instructional designers. While many teachers are able to identify some CWAs, many incorrect answers were missed by teachers while other answers that teachers suspected may be common were found to be less frequent in practice.

From our first analysis, we highlight that the definition of CWAs, determined by the chosen frequency threshold, may be further optimized to help bring greater attention to the most prominent errors made by students. While raising this threshold helps to identify the overall most common errors, this may also result in many errors being overlooked by teachers. Conversely, lowering the threshold may require teachers to spend more time providing individual feedback on more scarce errors instead of focusing on other instructional or tutoring methods that may be more effective. Furthermore, there is a limited understanding of CWAs and how to remedy them. Historical data on CWAs can play a pivotal role in answering various questions regarding CWAs. Do CWAs change over time? What are the factors that can drive changes in CWAs? How often should we be analyzing CWAs and generating CWAFs? Do certain types of feedback lead to better learning outcomes than other types of feedback? Future works exploring feedback could examine the effects of different features within the feedback messages and release guidelines for teachers and instructional designers on CWAFs.

Upon implementing these CWAFs, we observed that the feedback, on average, led to lower mastery and higher wheel-spinning among students working on mastery-based activities. If we factor in the negative effects of the CWAFs with the inaccuracy of teachers at predicting CWAs and the amount of time and effort that went into identifying and generating the CWAs and CWAFs, the approach of proactively identifying the CWA taken by the instructional designers of the two activities presented in this paper seems highly inefficient. Furthermore, the teachers also failed to identify several CWAs the students made while working on the problems, especially on the "2-Step Equations". Brown et al. [3] observed that students working on basic arithmetic problems can reach the same incorrect answer using different approaches, which required the tutors to

Table 8: Models Estimating Interactions Between Prior Performance and Common Wrong Answer Feedback (CWAF) Effects on Mastery by Activity

	Both Activities		2-Step Equations		Order of Operations	
Predictors	*Log-Odds*	*SE*	*Log-Odds*	*SE*	*Log-Odds*	*SE*
Intercept	3.34***	0.09	3.06***	0.09	4.29***	0.17
CWAFs (Treatment)	-0.17*	0.07	-0.18*	0.08	-0.06	0.18
Prior Problem Correct (Z-Score)	0.77***	0.04	0.80***	0.05	0.79***	0.10
Treatment X Prior Problem Correct	-0.11*	0.05	-0.80	0.06	-0.36*	0.14
Random Effects						
σ^2	3.29		3.29		3.29	
τ_{00}	1.62_t		1.45_t		2.03_t	
ICC	0.33		0.31		0.38	
N	564_t		443_c		191_c	
	21793_i		15835_i		5958_i	

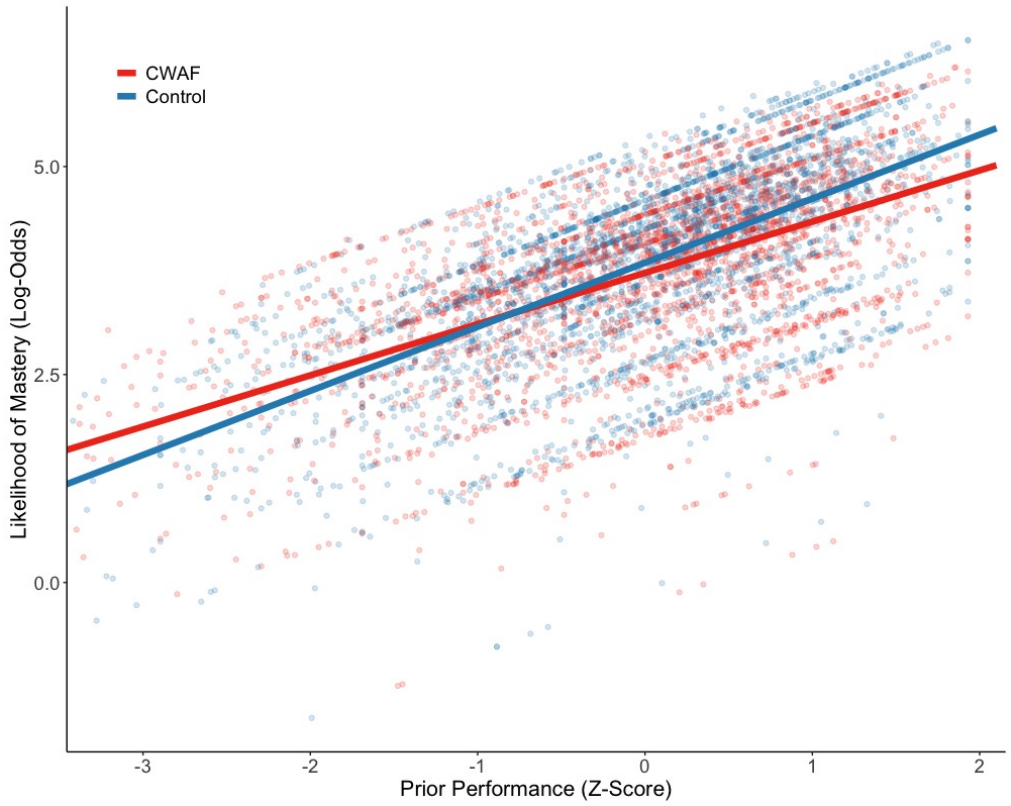

Figure 5: Interaction between students' prior percent correct and the predicted probability of mastery for the "Order of Operations activity" by condition.

first identify the wrong approach before providing the appropriate feedback. The process of identifying the students' approach was the primary factor in facilitating learning among students. It is our belief and recommendation that all future work exploring CWAs should leverage historical data when analyzing CWAs and generating CWAFs.

The study measuring the effectiveness of CWAFs adopts an intent-to-treat analysis where we examined the learning outcomes based on all students. We did not have information on the CWAs for the problems in the control condition as the problems were similar but not the same. As such, it is difficult to determine whether the effects we observe can be attributable to differences in the number

of common wrong answers experienced by students across the two conditions; given the large sample size of the study, this is likely to have little effect overall on our results, but can still be viewed as a limitation. Ideally, future studies could more accurately measure effects by comparing students who received CWAF in the treatment with students in the control group who *would have* received CWAF if they had been randomized to treatment.

It is also not clear from our current analyses whether students truly attended to the feedback they were given within the treatment condition. Recent work by Gurung and colleagues [11] utilized response time decomposition to identify students who are likely devoting attention and effort to tutoring and feedback they receive through the system. Student attention and consideration of feedback could be a large factor that mediates the overall effectiveness of CWAF. As prior work [26] found that learning gains were impacted by the length of the feedback, it may be the case that this attribute could also interact with a student's likelihood to read the CWAF; conversely, however, there is likely a trade-off in that shorter messages may be insufficient to provide students with enough information to effectively remedy the gap in knowledge. Similar to this, recognizing from other prior work [3] that different student errors may result in the same CWA, it is also possible that teachers authoring such feedback may misidentify the more prominent cause for the error. If the CWAF addresses an error that the student did not produce, it may cause greater confusion and ultimately cause students to lose trust or disengage with the system. Regardless, as it is found in our study that the CWAF was either ineffective or even negatively impact student learning, such a finding emphasizes a need to closely examine aspects of this feedback to understand what might be contributing to these outcomes.

We implore researchers in the domain of learning analytics to use our findings in this paper to explore the detection of CWA further and generate CWAFs to, with caution, explore the effectiveness of different feedback structures. At the same time, our findings in this paper indicate that CWAFs, on average, have a negative effect on student learning outcomes. Further analysis and additional research are required before the learning analytics community can reach an informed consensus on the effectiveness of CWAFs, given the counter-intuitive nature of this finding in light of other works recognizing the benefits of feedback for learning.

8 CONCLUSION

This paper presents additional evidence in line with prior work in the education domain, highlighting the nuanced challenges in identifying CWA and generating effective CWAFs that can remedy the various factors that resulted in the CWA. Our analysis underscored the risks of a proactive approach in identifying CWAs and generating the CWAFs as a large portion of the CWAFs that the teachers and instructional designers predicted were not made by the students. We also observed that CWAFs, on average, can lead to lower mastery and higher wheel-spinning amongst students–both undesired learning outcomes. Furthermore, we analyzed the personalization effects of CWAFs. While the effects were not significant, the data indicated that high-performing students were less likely to benefit from the CWAFs, resulting in lower mastery and higher wheel-spinning. While these findings add noteworthy value to the field of research exploring CWAs and the use of CWAFs in CALPs, researchers exploring CWAFs should not be discouraged by our findings. As mentioned in our discussion and recommendation sections, we believe that the learning analytics community will need to explore CWA and CWAFs further before we, as a community, can reach an informed opinion on CWAs and CWAFs.

ACKNOWLEDGMENTS

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] National Governors Association et al. 2010. Common core state standards. *Washington, DC* (2010).
[2] Joseph E Beck and Yue Gong. 2013. Wheel-spinning: Students who fail to master a skill. In *International conference on artificial intelligence in education*. Springer, 431–440.
[3] John Seely Brown and Richard R Burton. 1978. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive science* 2, 2 (1978), 155–192.
[4] John Seely Brown and Kurt VanLehn. 1980. Repair theory: A generative theory of bugs in procedural skills. *Cognitive science* 4, 4 (1980), 379–426.
[5] Richard R Burton. 1982. Diagnosing bugs in a simple procedural skill. *Intellinget Tutoring Systems* (1982), 157–184.
[6] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.
[7] Linda S Cox. 1975. Diagnosing and remediating systematic errors in addition and subtraction computations. *Arithmetic Teacher* 22, 2 (1975), 151–157.
[8] 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.
[9] JA Ferreira and AH Zwinderman. 2006. On the benjamini–hochberg method. *The Annals of Statistics* 34, 4 (2006), 1827–1849.
[10] 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.
[11] 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.
[12] 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.
[13] Heather C Hill and Mark Chin. 2018. Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Educational Research Journal* 55, 5 (2018), 1076–1112.
[14] Kim M. Kelly, Yan Wang, Tamisha Thompson, and Neil T. Heffernan. 2015. Defining Mastery: Knowledge Tracing Versus N- Consecutive Correct Responses. In *proceedings of the 8th International Conference on Educational Data Mining*. Association for Computing Machinery, New York, NY, USA, 39–46. http://web.wpi.edu/Pubs/ETD/Available/etd-041416-122623/unrestricted/wang.pdf#page=42
[15] Cheng-Fei Lai. 2012. Error Analysis in Mathematics. Technical Report# 1012. *Behavioral Research and Teaching* (2012).
[16] 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.
[17] Susanne Narciss. 2013. Designing and evaluating tutoring feedback strategies for digital learning. *Digital Education Review* 23 (2013), 7–26.
[18] Bobby Ojose. 2015. *Common misconceptions in mathematics: Strategies to correct them.* University Press of America.
[19] Bobby Ojose. 2015. Students' Misconceptions in Mathematics: Analysis of Remedies and What Research Says. *Ohio Journal of School Mathematics* 72 (2015).
[20] Ethan Prihar, Manaal Syed, Korinn Ostrow, Stacy Shaw, Adam Sales, and Neil Heffernan. 2022. Exploring Common Trends in Online Educational Experiments.

In *Proceedings of the 15th International Conference on Educational Data Mining*. 27.
[21] Sheryl J Rushton. 2018. Teaching and learning mathematics through error analysis. *Fields Mathematics Education Journal* 3, 1 (2018), 1–12.
[22] 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.
[23] 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.
[24] Raymund Sison and Masamichi Shimura. 1998. Student modeling and machine learning. *International Journal of Artificial Intelligence in Education (IJAIED)* 9 (1998), 128–158.
[25] Kurt VanLehn. 1982. Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. *The Journal of Mathematical Behavior* (1982).
[26] 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.
[27] John Woodward and Lisa Howard. 1994. The misconceptions of youth: Errors and their mathematical meaning. *Exceptional Children* 61, 2 (1994), 126.
[28] Richard M Young and Tim O'Shea. 1981. Errors in children's subtraction. *Cognitive Science* 5, 2 (1981), 153–177.