Understanding and Optimizing Mistake Messages in Mathematics

A Major Qualifying Project

submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

degree of Bachelor of Science

By:

Shannen Lin

Siddhartha Pradhan

Chayanne Sandoval-Williams

Date:

04/26/2023

Report Submitted to:

Professors Neil Heffernan and Stacy Shaw

Worcester Polytechnic Institute

Table of Contents

Acknowledgements

We would like to express our deepest appreciation and gratitude to Professor Neil Heffernan and Professor Stacy Shaw for taking time out of their busy schedules to provide invaluable insights and recommendations. They have been great mentors who have helped us grow as research scientists, in addition to being amazing advisors.

Many thanks to Ashish Gurung (PhD) and Kirk Vanacore (PhD) for their help in the intent to treat analysis for Study 1. Additionally we would like to thank Alena Egorova (PhD) for help with new methods of data analyzation for Study 2. Finally, we would like to thank Wesley Lo for his work on the IQP that was the catalyst for this project.

Abstract

With the advancement of technology, the field of education has also improved dramatically. For example, technology has provided teachers the ability to provide immediate feedback to students through the use of tools like ASSISTments, an online learning platform where teachers can assign homework to their students and view analytics regarding their students' progress. While these new technologies have improved learning, it also raises new questions. Specifically, we were interested in how different types of computer-provided feedback could affect students' performance and emotions. In addition, we also looked at math anxiety, which is increased nervousness and physiological reactivity when engaging with math, to see how it interacts with different types of feedback (Luttenberger et al., 2018). We conducted 2 studies, the first experiment evaluates the use of feedback and positive feedback messages to students who made common wrong answers. The second experiment analyzes the relationship between math anxiety and feedback tone on participants' emotions and attitudes towards math.

Understanding and Optimizing Mistake Messages in Mathematics

Math is necessary and important in many occupations, including high-earning careers related to Science, Technology, Engineering, and Math (STEM) fields. As such, math is often regarded as being fundamental to an individual's, as well as the nation's, economic success (Beilock & Maloney, 2015). The ability to think mathematically has also been shown to improve one's ability for reasoning, problem solving, and making connections (Stacey, 2006). While math does provide many benefits, it can be difficult for many people. Not only can the content itself be quite difficult due to its ties to logic and problem solving, but it can also cause undue stress and anxiety.

Math anxiety is defined as increased nervousness and physiological reactivity when engaging with math, such as solving a math problem or taking a math assessment (Luttenberger et al., 2018). Within an educational setting, it is very common to come across those with math anxiety. It is estimated in the United States that 25% of 4-year college students and 80% of community college students experience moderate to high math anxiety (Beilock & Maloney, 2015). In one survey, 93 percent of Americans indicated that they had experienced some level of math anxiety (Blazer, 2011). The effects of math anxiety include decreases in math achievement, but may also lead to people avoiding math or math-related occupations (Beilock & Maloney, 2015). Interestingly, having math anxiety is not necessarily a result of poor math ability. While some studies show that low math achievement is correlated with future math anxiety, there are other theories that suggest that it is the anxiety that disrupts important cognitive functioning needed to solve math problems, causing people to perform worse than their actual math abilities (Ramirez et al., 2018).

An important question for education researchers is how can we help to alleviate stress in these students with math anxiety? To answer these questions, we need to look at the causes of math anxiety. Researched causes of math anxiety include genetic predispositions and socio-environmental factors (Ramirez et al., 2018). Socio-environmental factors include one's parents and teachers, as well as one's own interpretation. One study by Maloney et al. (2015) found that with math-anxious parents, the more they helped their child with math homework, the higher the child's math anxiety after a year. Children who got homework help from math-anxious parents had higher math anxiety than children who got homework help from non-math-anxious parents. Similar results have been found with math-anxious teachers. A study by Beilock et al. (2010) found that girls taught by more math anxious female teachers showed worse math achievement and lower confidence in their math abilities than girls not taught by math anxious teachers. Given how teachers and parents often act as role models for children, it is understandable that children may adopt similar attitudes towards math as their role models. One of the more recent theories of math anxiety posits that it is how a child interprets these and other disfluent experiences in math that may explain why they may or may not go on to develop math anxiety (Ramirez et al., 2018). For example, if a student struggles to solve a math problem, they may internalize this as an innate lack of mathematical ability, leading to the development of math anxiety. Conversely, students who understand that math can be difficult and allow themselves to learn from failures are less likely to develop math anxiety. This interpretation of math anxiety suggests that by helping students reinterpret struggle or failure more positively, we can reduce math anxiety.

In discussing attitudes towards math in general, it is important to discuss the ideas of self-efficacy and growth mindset. Math self-efficacy is a person's perceived confidence in their math skills. A study done by Huang et al. (2019) looked at the effect of self-efficacy and math anxiety on STEM career interest. They had 152 middle school students take a survey that showed that self-efficacy was negatively correlated with math anxiety and positively correlated with STEM career interest. Having a growth mindset means that the person believes that by working on their skills, they will be able to improve over time. A fixed mindset refers to the opposite, believing that intelligence is innate and cannot be improved. A longitudinal study examining 1449 high school students over the span of a school year found that growth mindset beliefs predicted higher math achievement and higher STEM career aspirations (Degol, 2018). What these studies show is that one's mindset affects their attitudes towards math.

One central focus of early growth mindset research was helping students embrace failure and more productively interpret their struggle experiences (Diener & Dweck, 1978). Often this was accomplished through feedback made to students– either praise for intelligence or praise for effort (Mueller & Dweck, 1998). Indeed, feedback, defined as information given to a person in response to one's performance or understanding, has been shown to have a great influence on learning and achievement. Feedback can tell a person whether they were correct or incorrect, subtly shift student focus (as is the case with praise for effort growth mindset interventions), suggest alternative strategies, give explanations, or provide hints that can point students in the right direction (Hattie & Timperley, 2007).

Yeager et al. (2014) wanted to see what kinds of critical feedback were more beneficial to students, especially among minority adolescents. In their studies, they used a method called WISE feedback, in which the teacher gave critical feedback on an essay to the students that were designed to emphasize the teacher's high standards for the student and the teacher's belief that the student could meet those expectations. They found that students who received WISE feedback were more likely to submit a revision of their essay, and that the quality of their essay was also improved, demonstrating an effect of a feedback message on important student outcomes.

Another study looked at how different types of feedback could impact participants' motivation while playing a brain training game (Burgers et al., 2015). Two dimensions of feedback were looked at: valence (whether the feedback was positive or negative), and the feedback type (descriptive, comparative, evaluative). They found that evaluative feedback (telling the participant whether their performance was excellent or poor) increased future game play while comparative feedback (telling the participant whether their performance was faster or slower than their peers) decreased future game play. They also found that positive feedback boosted intrinsic motivation and perceived competence. Collectively, these studies suggest that well designed feedback messages can change student behaviors and attitudes. For students who face additional challenges, such as those with math anxiety, these feedback messages may hold even more importance as recent theories suggest they can help students more positively interpret struggle and failure (Ramirez et al., 2018).

Providing appropriate and helpful feedback was often quite difficult in the past. Within the classroom, it was not easy for a teacher to know when students were struggling in the class and how to help them. The only method to provide feedback to students was by having a teacher look over the student's work and then give verbal or written feedback. Obviously, in a class of many students, the feedback is often very delayed due to the teacher having to manually grade so many assignments, which often makes the feedback useless for struggling students.

The advancement of educational technology, a field that involves applying modern technology to improve the quality of teaching and learning, has resulted in the evolution of education in new ways, especially by improving how feedback is provided to students (Lazar, 2015). With advancements in technology, it is now possible to give computer-generated messages to students right when they answer a question incorrectly rather than waiting for the teacher's feedback later on. It is also possible to tailor those feedback messages to help students more positively interpret struggle and failure right as they are struggling rather than after the fact.

The use of computer-generated messages, however, brings up a new concern that has yet to be thoroughly researched. That is, with computer-generated messages, we do not know what types of messages might be better received by students. Depending on the tone and context, messages could come off to the student as encouraging, motivating, or even inadvertently patronizing.

The current study explored how mistake messages might be used and modified to boost student performance in mathematics. Specifically, we redesigned general mistake messages (feedback for incorrect answers) to offer more social and emotional support as well as redesigned the language to help students reinterpret their failure experiences. The goal of this study was to assess whether slight modifications to these messages might improve student performance in math. An additional goal of the study was to better assess the nuance of *how* students interpreted these messages, such as whether one's math anxiety level, may influence how they interpreted the same messages.

This study consisted of two studies. In the first study, we wanted to see if the use of feedback and positive feedback messages for common wrong answers would impact students' performance in a large online educational platform. In study 2, we conducted a laboratory experiment to look at how math anxiety and feedback tone when getting a question incorrect may influence participants' emotions and attitudes towards math.

Study 1

Introduction

The first study builds off of a previous study conducted on ASSISTments where researchers analyzed whether receiving a feedback message after making a mistake would influence student performance and behavior (Lo, 2021). In the previous study, students were exposed to one of two conditions: the control group which received no feedback other than correct/incorrect, and a treatment group that received enhanced feedback messages if the student made a common wrong answer. While it was predicted that enhanced feedback would improve student performance, as it displayed the likely step students made an error, there were no statistically significant differences in performance between the two groups. It was recommended that in future work, researchers could work toward improving the tone of these messages, as the feedback messages may have been interpreted as harsh or overly critical.

In the current study, students were randomly assigned into one of three conditions: no feedback (control), feedback (treatment 1), and feedback + messaging (treatment 2). Mistake messages were also redesigned in this study, students who were in the no feedback (control) condition were only told that their answer was wrong. Next, students who were assigned to the feedback (treatment 1) condition were shown general mathematical feedback. Finally, students in the feedback + messaging (treatment 2) condition were shown the feedback and additional messaging that was designed to help students reinterpret mistakes, struggles, or failures.

Research Questions

This experiment ran within ASSISTments and evaluated the use of feedback and positive feedback messages to students who made common wrong answers on the PSABTZFT problem set. The preregistration document can be found in Appendix A. In particular, this experiment explores whether providing feedback on common wrong answers improves student learning compared to no feedback.

We seek to explore the following questions 1) is there a difference in post-test correctness and mastery speed for students who received general mistake messages vs those who did not? and 2) was there an additional benefit of receiving a mistake message with feedback and messaging vs. a general one for next-problem correctness and mastery speed?

Hypothesis

We hypothesize that students assigned to feedback + messaging will show greater completion rates of the problem sets compared to feedback alone, and may show a benefit of next problem correctness, and mastery speed (getting three problems in a row correct). We believe both feedback conditions will show better completion, next-problem correctness, and mastery speed compared to the no feedback condition.

Method

Participants

Data from 2067 students were collected in the study, from which two students were dropped due to not being assigned a condition, presumably due to them not starting the assignment. The remaining 2065 students' assigned condition distribution can be seen in Table 1. There were slightly more students assigned to treatment 2. Initially, we considered the possibility that the system may have assigned multiple conditions to a student, but checking for duplicates in the assigned condition for students at the assignment level proved otherwise. We concluded that there may have been a slight issue with the randomization code in the system.

Treatment	First Action	Count	Total
Control	correct_response	329	627
	wrong_response	200	
	hint_requested	88	
	answer_requested	6	
	url_requested	4	
Treatment 1	correct_response	316	613
	wrong_response	202	
	hint_requested	92	
	url_requested	3	
Treatment 2	correct_response	359	677
	wrong_response	231	
	hint_requested	82	
	answer_requested	5	

Table 1: Count of first action and total students based on assigned condition

Procedure

In this study, we worked with data from ASSISTments, an online learning platform where teachers can assign homework to their students and view analytics in order to gain insight on how to improve their teaching. In addition, students are given immediate feedback by the computer while completing their assignments.

Students were randomly assigned into one of three conditions: no feedback (control), feedback (treatment 1), and feedback + messaging (treatment 2). Students who were in the no feedback (control) condition were only told that their answer was wrong. Next, students who were assigned to the feedback (treatment 1) condition were shown general mathematical feedback. Finally, students in the feedback + messaging (treatment 2) condition were shown the feedback and additional social emotional messaging that was designed to help students reinterpret mistakes, struggles, or failures. Sample mistake messages for each of these conditions can be found in Figure 1.

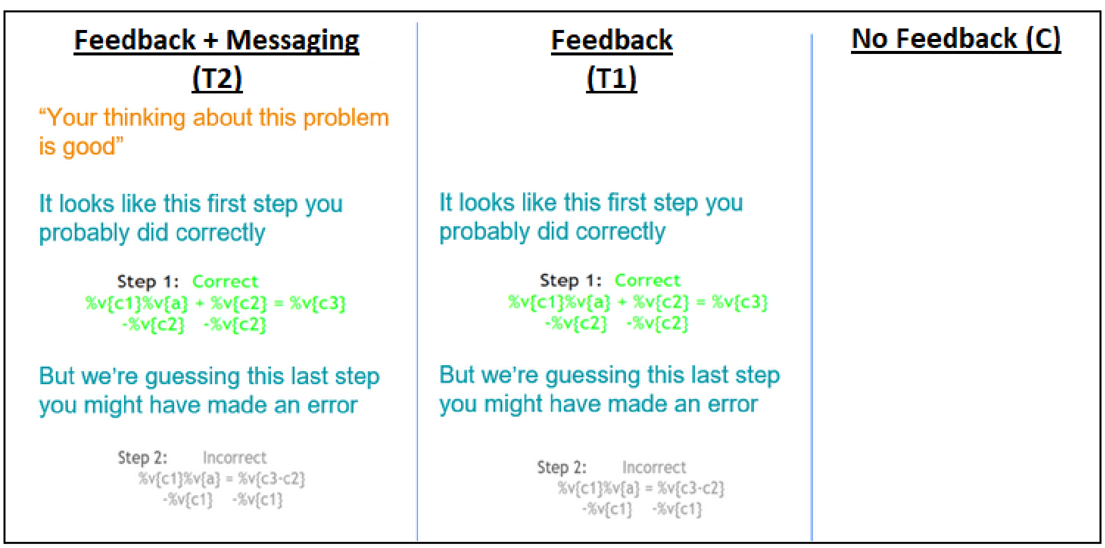

Figure 1: Sample mistake messages for Treatment 2, Treatment 1, and Control (left to right)

The mistake messages were only shown to students who entered an expected common wrong answer (ECWA), this included common mistakes students made while solving simple equations.

In order to complete the assignment a student would need to complete at least 5 problems. The first problem is the condition problem, in which the students are exposed to the mistake message given they enter an expected common wrong answer (ECWA). For additional randomization, the condition problem was selected from one of the three problems, with each problem having different ECWAs. The second problem that followed was an immediate post-test in which each condition received the same problem without any feedback. After completing the post-test questions the students were required to acquire mastery (i.e. consecutively getting three problems correct) of the topic. This problem flow of the experiment can be seen in Figure 2.

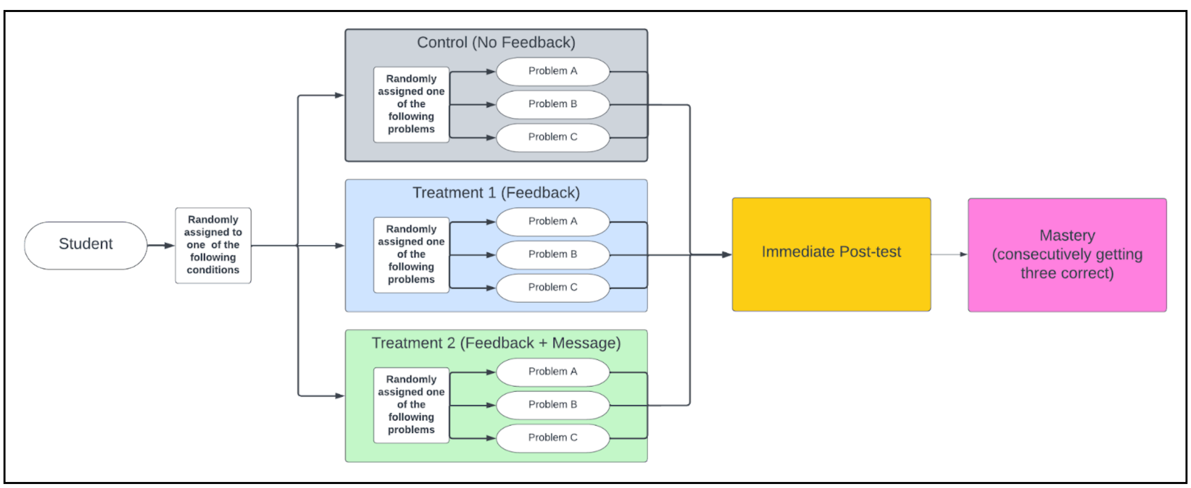

Figure 2: Problem flow for experiment

Next, in order to test our hypothesis, participants would have to be exposed to the treatment and complete the remaining problems. Firstly, we dropped students who were not assigned a condition. Next, since our hypothesis involved the time participants spent and correctness from the next problem, we dropped students who had not completed the post-test. Finally, we kept participants who made an ECWA in the first attempt for the condition problem. We only considered ECWA in the first attempt so that participants who performed other actions such as requesting hints were filtered out. The exact filtering process can be seen in Figure 3.

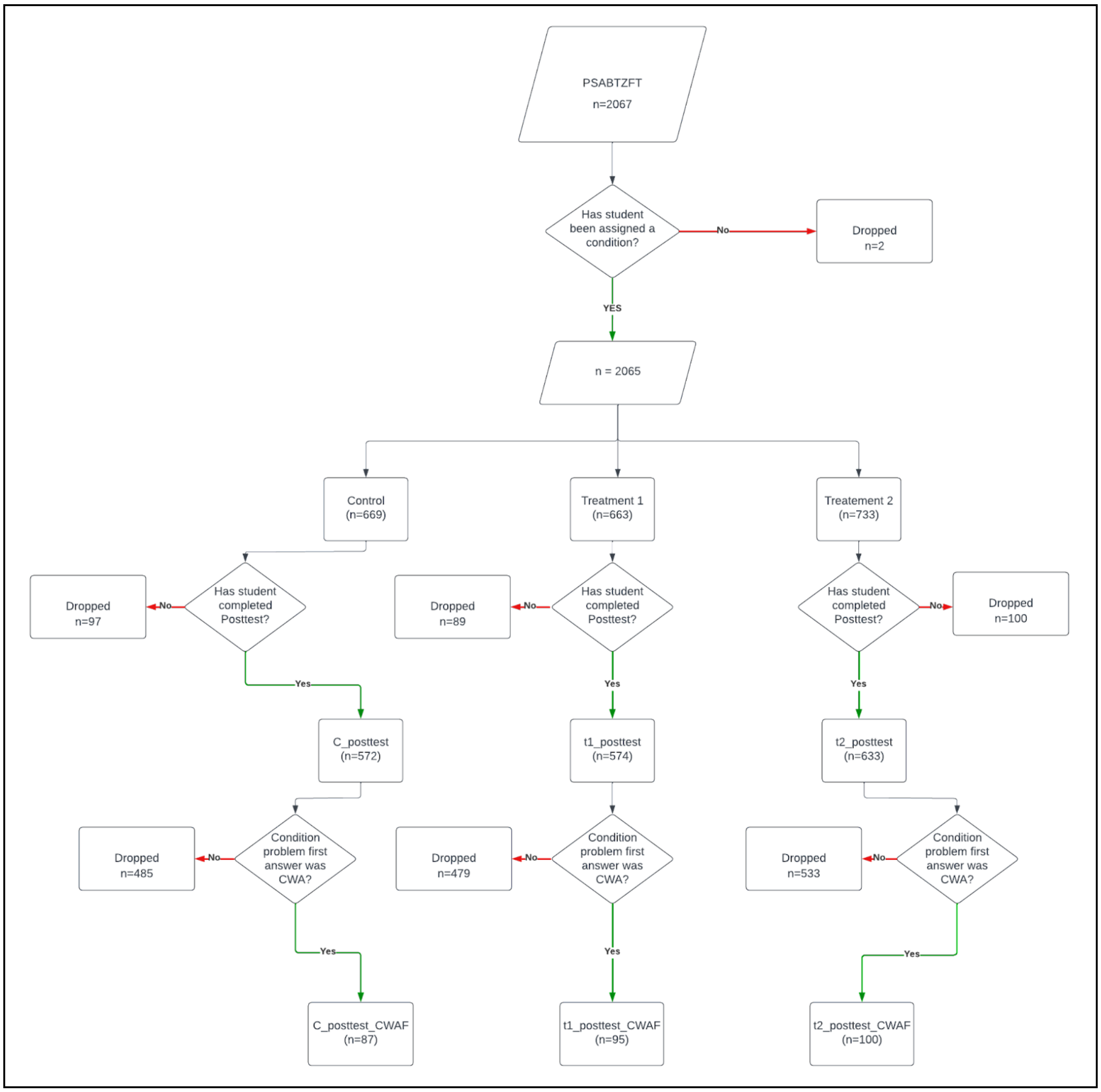

Figure 3: Filtering process for raw data

Measures

Next Problem Correctness. For participants who got the first condition problem wrong with an ECWA, and hence were exposed to the treatment, we recorded whether they got the next problem–in this case post test–correct or incorrect.

Mastery Speed. A calculated field that takes the inverse of the number of problems a participant attempted before correctly answering three problems in a row.

Time on Task (Next Problem). The total number of seconds a participant spent on the subsequent problem (post test) after being exposed to the treatment in the condition problem. We plan on the time students spent on a problem as a measure of persistence.

Rank Score. We ranked the participants values ranging from 3 to 0. The highest rank (3) is assigned to participants who got the immediate post-test and first question of the mastery section correct. The problem closest to the condition problem (i.e. immediate post test) was assigned a higher weight on determining ranking, as we expected the condition to have a higher effect on the problem that directly follows it rather than the mastery problems. The exact ranking assignments can be found in Table 2.

Assigned Rank Score	Answered Correctly		
	Condition Problem	Posttest Problem	Mastery 1st Problem
3	No	Yes	Yes
2	No	Yes	No
1	No	No	Yes
0	No	No	No

Table 2: Ranked score based on the questions answered correctly

Data Analysis Plan

In order to test the accuracy of any models we created and to prevent possible overfitting, we first sampled half of the participants in each of the conditions. This was achieved by dropping half the students based on whether their student ID was even or odd.

We planned on using the prior average correctness score of students as a covariate to potentially moderate the effect of our experiment. Prior average correctness is the aggregated average scores for students before the experiment started.

The distribution of the final analytical sample after cleaning and filtering is given in Table 3.

Treatment	First Action	Count	Total
Control	wrong_response	75	87
	hint_requested	12	
Treatment 1	wrong_response	78	95
	hint_requested	17	
Treatment 2	wrong_response	92	100
	hint_requested	8	

Table 3: Count of first action and total students based on assigned condition (filtered)

We planned to answer research question 1 (Control vs Treatment 2), with a Chi square test of independence, which will tell us if condition has a significant effect on post test correctness. To test our hypothesis that students in the Treatment 1 and 2 groups will show improved mastery speed and post test persistence compared to the students in the control group, we planned on creating linear models and running t-tests to check for statistically significant effects.

Mastery Speed. We plan on creating a linear model with mastery speed as our dependent variable and assigned condition as the predictor. Additionally, we planned on using the students' prior average correctness as a covariate to control for prior knowledge.

Time on task. We plan on creating a linear model with standardized post test time as our dependent variable and assigned condition as the predictor. Similar to mastery speed, we planned on using the students' prior average correctness as a covariate to control for prior knowledge.

Ranked Score. We plan on creating a linear model with ranked score as our dependent variable and assigned condition as the predictor. Similar to mastery speed and time on task, we planned on using the students' prior average correctness as a covariate to control for prior knowledge.

Results

To answer whether students assigned to treatment 2 showed higher post test correctness in comparison to the control group and if there was any additional benefit for students in treatment 2 in comparison to treatment 1, we performed a Chi-Square test of independence. A contingency table (frequency table) was created of participants who got the Post Test Item correct, by condition in Table 4. This Chi square test of independence will also answer research question 2 (Treatment 1 Vs Treatement 2), ie. is there an additional benefit of receiving a mistake message with feedback and messaging vs. a general one. The results of this test, X^2 *(2, N = 282) = 0.99, p > 0.05.*, suggests that there is no significant interaction between the assigned condition and posttest correctness.

Post Test Correct	**Control**	**Treatment 1**	**Treatment 2**	**Total**
Yes	46	50	53	**149**
No	41	45	47	**133**
Total	**87**	**95**	**100**	**282**

Table 4: Frequency Table of Participants who got the Post Test Item correct, by condition

To test the hypothesis that students in the treatment groups will show greater mastery speed than the control group, we first ran a two-tailed t-test with the means of each group. There was no significant effect for mastery speed on Treatment 2 Vs Control, *t(186) = 0.310045, p = 0.756975*. Similarly, there was no significant effect on mastery speed for Treatment 2 Vs Treatment 1, *t(194) = -0.0899913, p = 0.92841.* The descriptives for mastery speed of each of these groups can be seen in Table 5.

Condition	Control	Treatment 1	Treatment 2
Mean	0.199672	0.196654	0.197497
Std	0.0595034	0.0578552	0.0598838

Table 5: Descriptive statistics of Mastery speed based on assigned condition

Next, to test the hypothesis that students in the treatment groups will show greater persistence (i.e., post test time on task than the control group), we first ran a linear model with assigned condition as predictor. There was no significant effect of condition on the standardized post test time on task, $R^2 = .005$, $F(2, 279) = 0.687$, $p = .504$. The full results for the linear model can be found in Appendix B.

Finally, to test the hypothesis that students in the treatment groups will show higher ranked scores than the control group, we first ran a two-tailed t-test with the means of each group. There was no significant effect on ranked score for Treatment 2 Vs Control, $t(186) = -0.778256$, $p = 0.437513$. Similarly, there was no significant effect on ranked score for Treatment 2 Vs Treatment 1, $t(194) = -0.940583$, $p = 0.348172$. The descriptives for ranked score of each of these groups can be seen in Table 6.

Condition	Control	Treatment 1	Treatment 2
Mean	1.778	1.755	1.921
Std	1.265	1.242	1.140

Table 6: Descriptive statistics of Mastery speed based on assigned condition

Study 1 Discussion

As our results suggested that the assigned condition, i.e., type of mistake message, had no effect on students' performance, a remaining question was whether this was driven based on differences in how students interpreted these messages. Some students may be encouraged by a message with a positive tone, while others may find it patronizing. These interpretations can change the way that students react to the mistake messages. For the same message, some students might respond positively and some might not, washing out an overall effect of the assigned condition. Thus, we next asked if there was variation in how students interpret these messages? Does this depend on their math anxiety? Since the participants in study 1 were middle schoolers, we cannot ask them specific questions (such as math anxiety or their responses to the mistake messages) that take them out of the role of a student and into the role of the participant. So in study 2, we tested this hypothesis.

Study 2

Introduction

Following Study 1, we were interested in seeing how different people would perceive different styles of messages. Would something that we considered to be an encouraging positive message be interpreted differently? What messages would end up being the most helpful? In Study 2, we wanted to separate the messages from the math to better understand what would work for the student.

By better understanding what a language or factors might inhibit or result in success for students, we hope to provide future suggestions for implementation into ASSISTments and other educational technology.

We developed a study in the laboratory to test WPI undergraduate students here at WPI. Being able to break down the barrier between student testing and student feedback, we hoped to figure out what information would best suit a student towards further success. In the creation of the study, we hoped to answer 3 important questions:

1. Do students respond more positively if they get positive mistake messages?
2. Do students perform better if they get positive mistake messages?
3. Does math anxiety play a role in how students interpret different messages?

Method

Participants

57 undergraduate students at WPI participated in the survey. They were recruited through an undergraduate research participation pool. Their participation in this pool provides them course credit in psychology courses. The sample was made up of 16 students who were Computer Science majors, 11 Biology/Biotechnology majors, 6 are Interactive Media and Game Design majors, and 4 were Psychology. The other students ranged from majors like Business to Data Science.

From the participants that took our study, 27 of them were female identifying, 22 were male identifying and 8 identified in some form outside of the gender binary.

Our final sample size was 57 participants before cleaning the data. When cleaning, we had to remove 1 participant that did not finish the survey, and 1 other participant that did not get any questions wrong, thus did not receive any mistake messages.

Measures

Math Anxiety. Math anxiety was measured using the Abbreviated Math Anxiety Scale (Hopko, Mahadevan, Bare, & Hunt, 2003). Participants were presented with nine items and asked to rate how anxious each item would make them feel (e.g., starting a new chapter in a math book, taking an examination in a math course, etc.) on a scale from 1 (low anxiety) to 5 (high anxiety). Scores across these 9 items and then combined to create a composite score of math anxiety.

For our math anxiety variable, we divided participants into those with low anxiety and those with high anxiety. This was done by calculating an average math anxiety score for each participant. Those who had a math anxiety score of less than 3 were placed in the low anxiety group while those with an anxiety score of more than 3 were placed in the high anxiety group. A value of 3 was chosen as the cutoff point as it is the median score for the AMAS (which ranges from 1 to 5).

Math Task. The math task presented in this study consisted of fourteen multiple choice questions taken from GRE sample tests. The topics of these questions ranged from algebra, geometry, and arithmetic. For each question, participants were given 90 seconds to answer, with a countdown timer displayed in the top left corner. Submitting an answer or failing to submit an answer within the time limit would take the participant to the next screen showing whether or not they answered the question correctly as well as displaying a mistake message.

Mistake Message Affect. The independent variable for this study was the tone of the mistake messages. In one condition, the tone of the mistake messages presented to participants were very positive, such as “You got it wrong, but that's okay, you did your best!” and “Your answer was incorrect; don't worry! This problem is hard for most people!”. In the other condition, the tone of the mistake messages were more neutral, such as “Your answer was not correct” or “Wrong. Make sure to read the question thoroughly.” The mistake messages can be found in Appendix D. For each mistake message they encountered, participants were asked to select emotions that the mistake message made them feel. These emotions ranged from positive emotions such as encouraged, motivated, or confident to negative emotions such as annoyed, belittled, or sad.

Emotional Response towards Mistake Messages. Participants were given a series of Likert statements related to their attitudes towards the mistake messages they saw during the math task. Some example statements include "The mistake messages made me stressed" and "The mistake messages were helpful". The scale items can be found in Appendix E. They were then asked to rate how much they agreed with each statement on a scale from 1 (strongly disagree) to 7 (strongly agree).

Mistake Message Preference. Participants were also asked to rate their preference for a list of sample mistake messages on a scale from 1 (not preferred) to 7 (highly preferred). The mistake messages were a mix of both positive and neutral tone messages.

Demographic Information. At the end of the study, participants were asked to report demographic information such as age, gender, and race/ethnicity. Participants also reported which math courses they had taken from a list of courses (algebra 1-2, calculus 1-4, probability, statistics, etc) as well as their college major to potentially account for any differences in math background.

Procedure

Participants entered the lab and sat in front of a computer. They were then instructed to sign the general consent form and the COVID-19 consent form, as guided by the IRB. Then they completed the Abbreviated Math Anxiety Scale, before completing the math task itself.

The math task was a quiz made up of 14 sample GRE Math questions. These varied in difficulty and were presented to the participants in a randomized order to better our chances of them receiving a mistake message.

Participants were randomly placed into one of two conditions, either positive or neutral. For those in the positive condition, after getting a question wrong, they might receive a mistake message like: “You almost got it; nice try!”. While those in the neutral condition may receive something like “That is NOT the answer”. For a full list of the mistake messages used in the study, please view the Appendix. These conditions did not change anything they saw besides the mistake messages they saw after getting a question wrong. Participants were unaware of these conditions until the debriefing in order to get an honest response.

Following the math task, participants were asked to answer questions about the mistake messages they were presented. They were asked how they felt about the specific messages they saw during their quiz.

Finally we asked every participant simple demographics questions to gain a sense of who was participating in our study. The demographics survey asked for participants simple information like age, major, gender identity, and ethnic background.

Results

The first question we sought to answer was whether students responded more positively if they received positive mistake messages. To examine this, we looked at how the participants rated the mistake messages they saw during the math task afterwards. Negatively valenced statements assigned a negative weight and the sum of the likert statements was used as a measure for emotional response. The value for positive emotional responses would be more positive and higher than the value of negative emotional responses. Linear regression predicting emotional response from assigned condition showed that the condition did have a statistically significant effect on emotional response ($b = 5.55$, $t(53) = 3.24$, $p = 0.002$) (see Figure 4). Those in the positive tone condition (M = 19.14, SD = 6.74) did have a more positive emotional response than those in the neutral tone condition (M = 13.59, SD = 5.99).

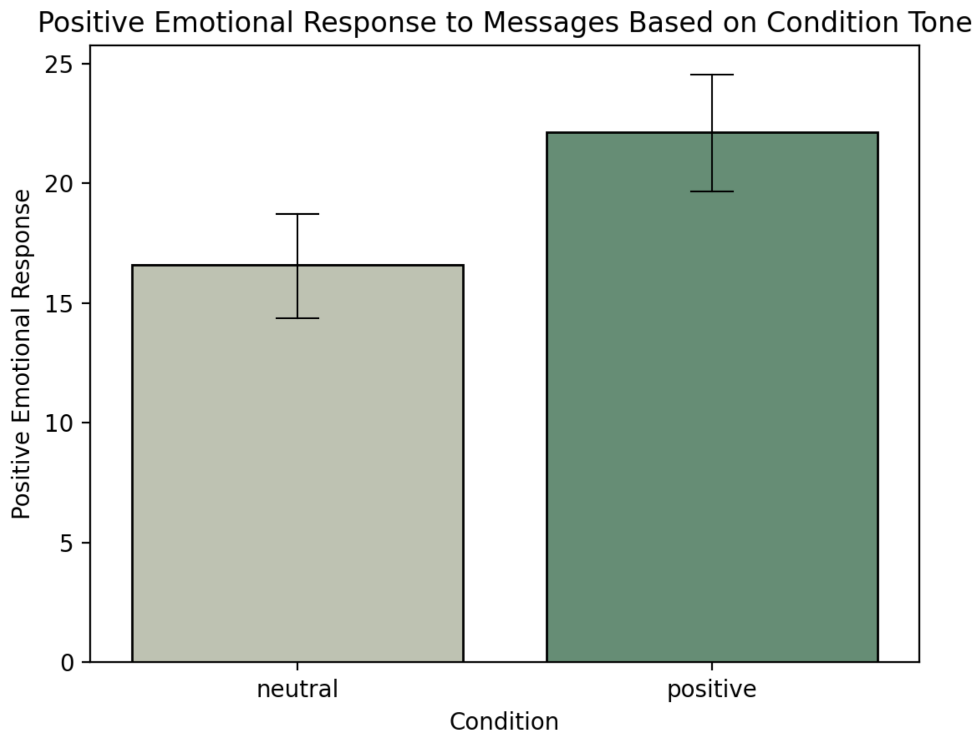

Figure 4: The effect of condition tone on participants' emotional response to the messages

To see whether math anxiety affected participants' attitudes towards the mistake messages, we looked at their preference ratings for a series of both positive and neutral mistake messages. For each participant, we calculated a total score for positive messages and a total score for neutral messages. We then conducted an independent t-test for each based on math anxiety level. We found that for positive tone messages, those with high anxiety (M = 4.82, SD = 1.41) showed a higher preference than those with low anxiety (M = 3.83, SD = 1.62), $t(55) = -2.28$, $p = 0.026$, two-tailed test (see Figure 5). For neutral tone messages, we saw an opposite effect where those with low anxiety (M = 4.42, SD = 1.10) showed a higher preference than those with high anxiety (M = 3.29, SD = 1.46), $t(55) = 3.25$, $p = 0.002$, two-tailed test (see Figure 5).

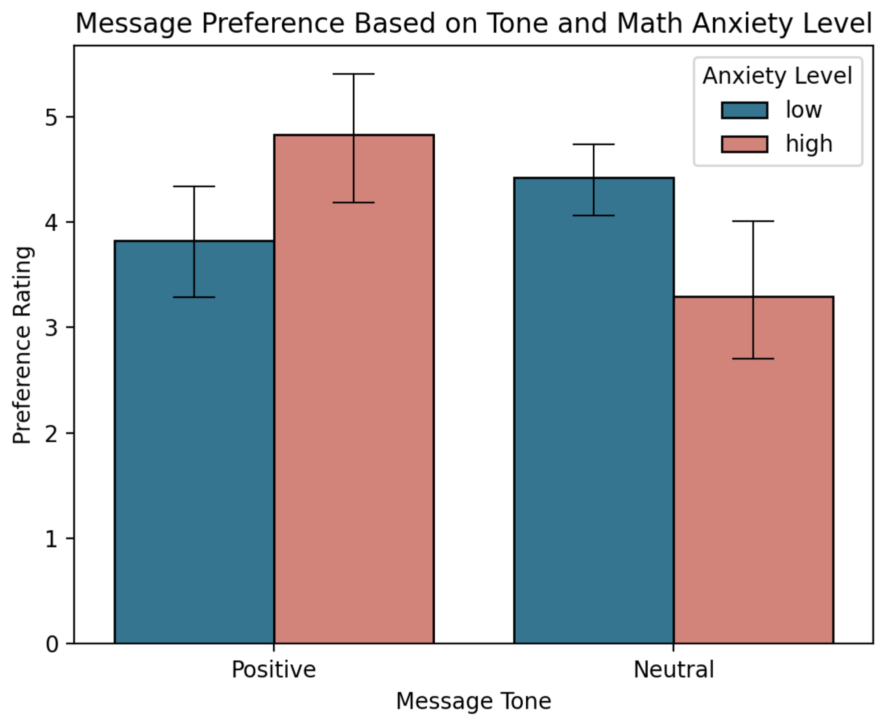

Figure 5: The effect of math anxiety level on the preference ratings for both positive and neutral messages

The second question we sought to answer was whether students performed better if they got positive mistake messages. In order to answer this question we performed a two-way ANOVA to analyze the effect of condition and math anxiety level on the participants' scores on the math task. To calculate each participant's score on the math task, we simply counted up the number of questions that the participant answered correctly. The ANOVA results showed that math anxiety did have a statistically significant effect on the math task score ($F(1, 53) = 4.916$, p

= 0.031) (see Figure 6). Those with low anxiety (M = 7.53, SD = 2.55) scored higher on the math task than those with high anxiety (M = 6.05, SD = 2.20). However, conditions did not have a statistically significant effect on final score ($F(1, 53) = 0.56$, $p = 0.456$) (see Figure 7). We also did not find any significant interaction between math anxiety level and condition tone on the math task score ($F(1, 53) = 0.551$ $p = 0.461$) (see Figure 8).

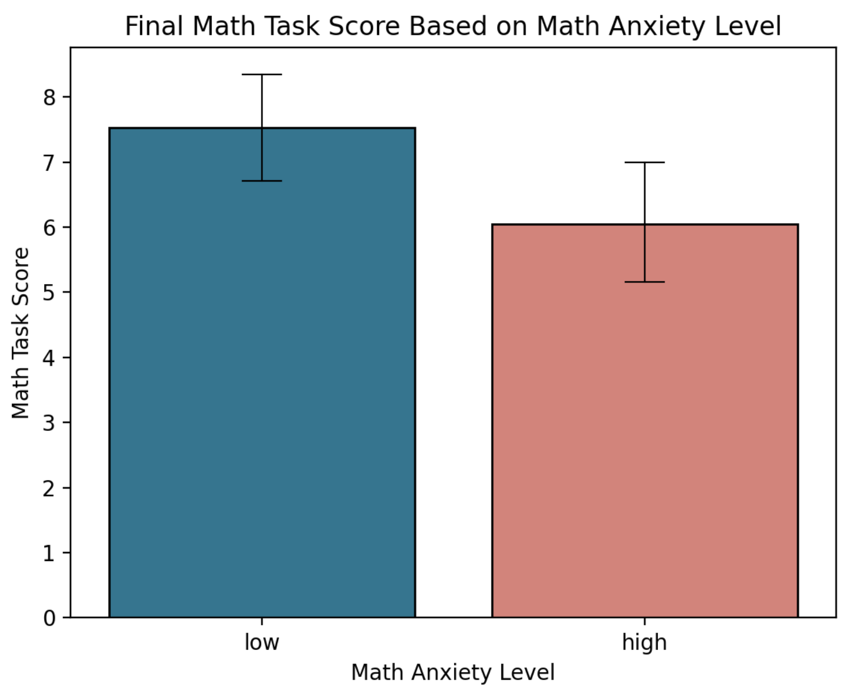

Figure 6: The effect of math anxiety level on the math task score

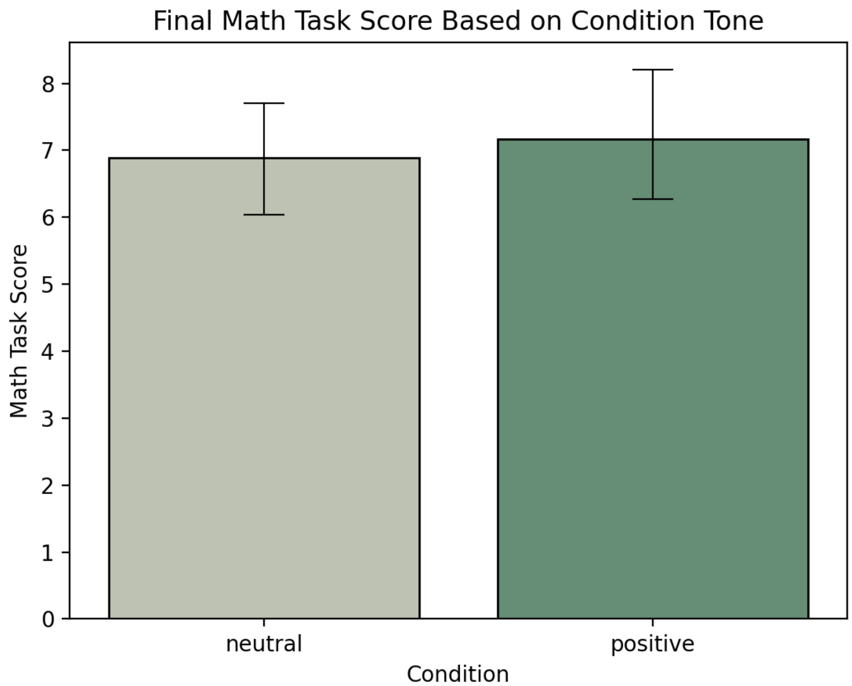

Figure 7: The effect of condition tone on the math task score

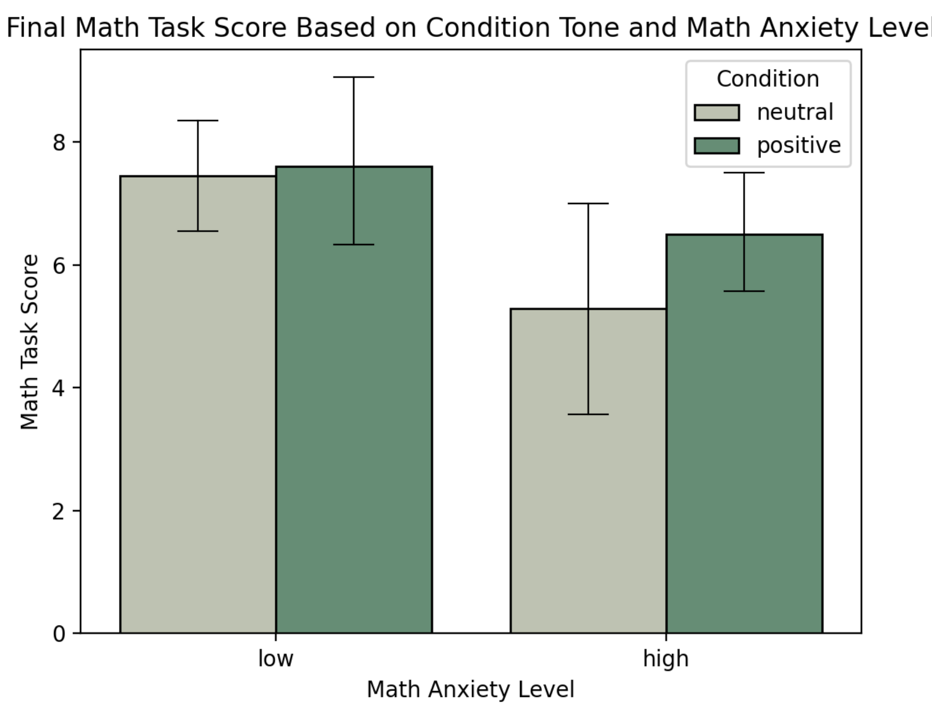

Figure 8: The effect of condition tone and math anxiety level on the math task score

Study 2 Discussion

Study 2 demonstrated that there are differences in how students interpret mistake messages. One factor was condition, seeing if the types of mistake messages affected students; the other was math anxiety. This showed that assigned condition had more of an effect on emotional response than it did performance. On the other hand, math anxiety had an effect on both performance, something backed up by previous literature, and message preference, a metric new to this study. The interaction between both condition and math anxiety didn't prove to be statistically significant for any metric. The results provided insights on how students prefer their feedback, and how additional factors like math anxiety can influence student learning.

Full Discussion

Across two studies, this MQP examined the efficacy of mistake messages in mathematics. Study 1 focused on K12 students and whether there were changes in performance based on the assigned condition. Study 1 did not find any statistical differences in post test correctness and other metrics based on the assigned condition. One reason our results differed from our hypothesis could be due to students not reading the mistake messages. Students are only shown the mistake message once, in the condition problem, after which the students are not exposed to the conditions, this could lead to an increase in the variation of post test correctness and other similar metrics. Another reason our results differed from our hypothesis could be that the mistake messages affected the students attitude to the assignment, however this change was not reflected in our measures. Another reason that we think is worth exploring is the variability in the interpretation of the message itself. Interpretation seems to be impactful in the classroom (Ramirez et al., 2018; Jamieson et al., 2012) and even some positive messages can be interpreted negatively, as it can feel condescending or as it can signal implicit bias (Harper et al., 2012).

Study 2 followed up on Study 1, working to understand both interpretation of the given mistake messages, but also preference towards messages and message tone. We found that the tone condition did affect the participants' attitudes towards the messages. Specifically, those in the positive condition felt a more positive emotional response to the messages they saw than those in the neutral condition. This is understandable as the mistake messages in the positive tone condition were meant to be more encouraging and promote more positive feelings.

For the math task score, we found that those with high anxiety did worse on the math task than those with low anxiety, which supports the results found by many previous studies

(Luttenberger et al., 2018). However, we found that condition tone did not have an effect on one's math task score, nor did it have any interaction with math anxiety. This is understandable considering that the amount of time that participants were exposed to the treatment was very short, so the possible effects of condition tone on one's performance might not be obvious.

One interesting finding was that those with high math anxiety tended to prefer positive tone messages more than those with low math anxiety. On the other hand, those with low math anxiety tended to prefer the neutral tone messages more than those with high math anxiety. The difference in interpretation between low and high math anxiety is quite interesting. Perhaps those with high math anxiety viewed the positive mistake messages as being encouraging. However those with low math anxiety, who also have greater confidence in their math skills, may have seen those same positive mistake messages as being condescending or patronizing. On the other hand, those with high anxiety are more likely to have viewed neutral tone messages, where the tone is ambiguous, as being more critical towards them. The results suggest that while mistake messages with a positive tone are overall shown to elicit a more positive emotional response compared to neutral messages, there do appear to be individual differences that may influence how a person interprets a mistake message.

The results of these studies tie back in with previous studies that have looked at how feedback can influence student learning. The research on feedback by Yeager et al. (2014) and Burgers et al. (2015) showed that feedback messages can be changed to influence students' attitudes and performance. The results from our study set the foundation for future research on how to design mistake messages to help students more positively interpret struggle and failure.

Limitations

The results of this study provide important information into further understanding how the tone of mistake messages can influence attitudes and performance in students. However, there are quite a few limitations in the study that should be highlighted.

In Study 1, a limitation was that students only ever saw a mistake message once. As the treatment was only shown once, the effect the treatment had on the participant could have been too small to detect.

There are also some limitations with Study 2 to discuss. Firstly, the sample size of our study was very small with only 57 participants. As our study divides participants by tone condition and math anxiety level, this means that the number of participants for a specific group was actually quite small. This made it difficult to obtain reliable results from our analysis due to the small size of each group. Future studies would benefit from having more participants.

Another limitation comes from the fact that this study was conducted in a lab. Due to this, we cannot be certain that the results of this study can be replicated outside of the lab setting. Conducting the study on a platform like ASSISTments would have provided ecological validity to our results, as we would be conducting the study with students actually working on homework assignments. However it would have been very difficult to accomplish as it would require a lot more time to gain IRB approval for such a large-scale study, which we unfortunately did not have. In Study 1 of the study, IRB approval was not necessary as they were only collecting data on each student's performance on the homework. On the other hand, Study 2 asked more in depth questions towards the participants, and would require IRB approval as well as each student taking part in the study to provide consent. Not only would this be quite the hassle to accomplish, it might be disruptful to students who are just trying to finish their homework.

There were some limitations that were caused by issues in the Qualtrics survey we created. One limitation was that the way in which we collected the data for mistake message affect, by having users select checkboxes for which emotions each mistake message made them feel, actually made it very difficult to analyze the data later on in a reliable way. It would probably be best in the future to use likert questions to determine the specific emotions that participants felt towards the mistake messages.

Another issue with our survey was that each mistake message was tied to one specific question rather than randomly shown to the participant. That means that mistake messages that were tied to more difficult questions were more likely to be seen by participants than mistake messages that were tied to easier questions. As a result, we were really only able to analyze the data for the mistake messages tied to the difficult problems. Future studies should write the survey so as to guarantee that each mistake message can be seen an equal number of times.

Finally, there was an error in the survey question looking at mistake message preference where those in the positive condition saw the likert statements on a 7 point scale while those in the neutral condition saw the likert statements on a 10 point scale. This required us to scale the scores in the neutral condition to match a 7 point scale. It is possible that this may have influenced the results in some way.

Recommendations for Future Work

In the future, it would be worth exploring more into the next problem correctness metric that was provided in Study 1 to better understand the correlation between performance and message type. Additionally, we believe that future work would benefit from taking the laboratory portion of the study into the real world, getting a chance to test it on students that are closer to the age of those using the ASSISTments system. This would provide more ecological validity for the results. With all of this in mind, there is potential for a team in the future to create an algorithm that takes in the factors of how a student learns, what they prefer and how math anxious they are to generate messages that better enhance their learning.

There is also a lot of potential to create additional software elements that can create a more personalized experience for students. Using factors that we studied here, like math anxiety and performance, along with other factors that we didn't study, like geographical educational differences and feedback, to create an algorithm that can personalize the feedback that students receive.

While we started this project believing the work we did would be finished as our team ended, we have come to realize we have only touched the surface of what can be done with this research. There is much more to learn and be done, all in an effort to continue to improve the learning experience for students.

Conclusion

As the education sector increasingly adopts educational technology, engineers, scientists, and computer scientists should collaborate towards designing systems that cater to the diverse learning needs of students. Pedagogy is not uniform across students and to enable optimal learning outcomes, personalized assistance akin to that offered by teachers in a traditional classroom setting is often required. Therefore, technological advancements that facilitate teacher-like, personalized feedback can significantly enhance the success of students. In this regard, seemingly minor features such as mistake messages and feedback may wield a considerable impact on students' educational progress.

Bibliography

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences*, *107*(5), 1860-1863.

Beilock, S. L., & Maloney, E. A. (2015). Math anxiety: A factor in math achievement not to be ignored. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 4-12.

Blazer, C. (2011). Strategies for Reducing Math Anxiety. Information Capsule. (ED536509). ERIC. https://files.eric.ed.gov/fulltext/ED536509.pdf

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, *48*, 94-103.

Degol, J. L., Wang, M. T., Zhang, Y., & Allerton, J. (2018). Do growth mindsets in math benefit females? Identifying pathways between gender, mindset, and motivation. *Journal of youth and adolescence*, *47*, 976-990.

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. Journal of Personality and Social Psychology, 36(5), 451–462.

Harber, K. D., Gorman, J. L., Gengaro, F. P., Butisingh, S., Tsang, W., & Ouellette, R. (2012). Students' race and teachers' social support affect the positive feedback bias in public schools. Journal of Educational Psychology, 104(4), 1149–1161. https://doi.org/10.1037/a0028110

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. http://www.jstor.org/stable/4624888

Huang, X., Zhang, J., & Hudson, L. (2019). Impact of math self-efficacy, math anxiety, and growth mindset on math and science career interest for middle school students: The gender moderating effect. *European Journal of Psychology of Education*, *34*, 621-640.

Jamieson, J. P., Nock, M. K., & Mendes, W. B. (2012). Mind over matter: reappraising arousal improves cardiovascular and cognitive responses to stress. Journal of experimental psychology: General, 141(3), 417.

Lazar, S. (2015). The importance of educational technology in teaching. *International Journal of Cognitive Research in Science, Engineering and Education*, *3*(1), 111-114.

Luttenberger, S., Wimmer, S., & Paechter, M. (2018). Spotlight on math anxiety. *Psychology research and behavior management*, 311-322. https://doi.org/10.2147/PRBM.S141421

Maloney, E. A., Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2015). Intergenerational effects of parents' math anxiety on children's math achievement and anxiety. Psychological Science, 26, 1480–1488. doi:10.1177/0956797615592630

Mueller, C.M., & Dweck, C.S. (1998). Praise for intelligence can undermine children's motivation and performance. Journal of Personality and Social Psychology, 75(1), 33-52.

Ramirez, G., Shaw, S. T., & Maloney, E. A. (2018). Math anxiety: Past research, promising interventions, and a new interpretation framework. *Educational psychologist*, *53*(3), 145-164.

Stacey, K. (2006). What is mathematical thinking and why is it important.

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., ... & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, *143*(2), 804.

Appendix A

Study 1 Preregistration Document

This is a preregistration for an experiment run within ASSISTments that evaluates the use of feedback and positive feedback message to students who make common wrong answers on problems.

Research Questions: Does providing feedback on common wrong answers improve student learning compared to no feedback? Does adding language designed to help students reinterpret struggle/failure to general feedback improve student learning or persistence compared to feedback without the language?

RI: Feedback vs. No feedback
R2: Feedback + Messaging vs. Feedback.

Data Collection: Data have been collected at the time of this preregistration but none of the researchers have looked at it and we currently do not have access to it.

Methodology:

Students who made a common wrong answer were randomly assigned into one of three conditions at the problem level

- No feedback: Students who were in this condition were just told their answer was wrong with no additional information about how it was wrong.
- Feedback: Students in this condition were shown general mathematical feedback.
- Feedback + messaging: Students in this condition were shown the feedback plus additional messaging before the feedback message designed to help students reinterpret mistakes/struggle/failure to solve

Content:
The actual content of the experiment is in the problem set PSABTZFT.
This exact content can be experienced by going to this link:
https://wvwu.assistments.org/public preview/link/cHJldmlldz10cnVlJmFub25BbGxvd2VkPXRyd WUmd2hpdGVMYWJlbGVkPWZhbHNlJnBwdXJsVmVyPTImYXNzaWdubWVudEIEPVBT MTQy MTQ5MCZvbkV4aXQ9aHROcHM6Ly93d3cuYXNzaXNObWVudHMub3Jn. As it's randomized each time, you will have to play a few times before you see all three conditions (it

does not log, so you don't have to worry about messing up our experiment) . This may be hard to "play" because you have to be able to guess what the common wrong answers are, which is when the feedback messages are generated. You can see the full problem set for the study here: https://drive.google.com/drive/folders/1rVlUgZe0Qie PYszFZOuOuJXad4frUUX?usp=sharing

Hypotheses:

We predict that students assigned to feedback + messaging will show greater completion rates of the problem sets compared to feedback alone, and may show a benefit of next problem correctness, and mastery speed (getting three problems in a row correct). We believe both feedback conditions will show better completion, next problem correctness, and mastery speed compared to the no feedback condition.

Data analysis plan:

We will begin by requesting half of the data of the experiment from the associated edu tech foundation. Using this half of the data, we will analyze the experiment with traditional outcomes used in the past, including Next problem correctness, Mastery speed (calculated by taking the inverse of the total number of problems that students had attempted by the time that they correctly answered three problems in a row) Completion of a problem set (Time spent on problem /subsequent problems).

Appendix B

Results for linear model with Standardized Post Test time as dependent variable

```
                            OLS Regression Results
==========================================================================================
Dep. Variable:     standard_posttest_time_on_task   R-squared:                       0.005
Model:                                        OLS   Adj. R-squared:                 -0.002
Method:                             Least Squares   F-statistic:                    0.6870
Date:                            Sun, 23 Apr 2023   Prob (F-statistic):              0.504
Time:                                    14:37:02   Log-Likelihood:                 398.71
No. Observations:                             282   AIC:                            -791.4
Df Residuals:                                 279   BIC:                            -780.5
Df Model:                                       2
Covariance Type:                        nonrobust
=======================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
Intercept                            0.0021      0.006      0.325      0.745      -0.010       0.015
assigned_condition[T.Treatment 1]    0.0078      0.009      0.883      0.378      -0.010       0.025
assigned_condition[T.Treatment 2]    0.0097      0.009      1.121      0.263      -0.007       0.027
==============================================================================
Omnibus:                      534.337   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           255049.774
Skew:                          11.320   Prob(JB):                         0.00
Kurtosis:                     148.581   Cond. No.                         3.86
==============================================================================
```

Note. The module and output was created by using the Python module Statsmodel.api

Appendix C

Sample GRE questions from the math task

In a series $x_1 + x_2 + x_3 \ldots x_n$ each of the terms is either 9 or 99. If the sum of the series is 531, then what could be the value of n?

- ○ 23
- ○ 26
- ○ 28
- ○ 29
- ○ 31

On a test consisting of 80 questions, Eve answered 75% of the first 60 questions correctly. What percent of the other 20 questions does she need to answer correctly for her grade on the entire exam to be 80%?

- ○ 85%
- ○ 87.5%
- ○ 90%
- ○ 95%
- ○ 100%

If a + b = 3(c + d), which of the following is the average(arithmetic mean) of a, b, c, and d?

- ○ $\frac{c+d}{4}$
- ○ $\frac{3(c+d)}{8}$
- ○ $\frac{c+d}{2}$
- ○ $\frac{3(c+d)}{4}$
- ○ $c + d$

Appendix D

List of mistake messages based on condition

List of Positive Mistake Messages that could be seen by participants during the math task:

- You were close, but not right yet!
- Keep trying, maybe you'll get the next question right!
- You almost got it; nice try!
- Not right, but maybe you'll get the next one!
- You got it wrong, but you gave it your best shot!
- This is incorrect, but don't worry!
- Whoops, almost there!
- This is wrong, but your effort is showing. Keep trying on the next questions!
- You got it wrong, but you are trying your best!
- Your answer was incorrect; don't worry! This problem is hard for most people!
- You got this one wrong, but don't give up yet!
- You got it wrong, but that's okay, you did your best!
- You got it wrong, but that's okay, most people do!
- You got it wrong, but that's okay, most people do!

List of Neutral Mistake Messages that could be seen by participants during the math task:

- INCORRECT
- No, that is not right
- This is not the correct answer.
- That is NOT the answer

- Sorry but your answer is wrong.
- Incorrect. Keep up your effort.
- Wrong.
- Your answer was not correct.
- Wrong. Make sure to read the question thoroughly.
- This is not the correct answer.
- That was not right. Are you sure you tried your best?
- That was definitely not the right answer.
- Your answer was wrong.
- Wrong. Try harder next time.

Appendix E

Likert questions asked during the experiment

List of L1 Likert questions measuring attitudes towards messages seen during the math task:

- Completing this task was stressful.
- The mistake messages were helpful.
- The mistake messages made me stressed.
- I would like to see more mistake messages like the ones in the assessment.
- The mistake messages inhibited my ability to move forward successfully.
- The mistake messages made me feel worse about my math ability.

List of L3 Likert questions measuring preferences towards a series of sample mistake messages:

- This is incorrect. Try again
- Don’t worry! Lots of students make a mistake on this problem.
- You almost got it! You’ll get it next time
- Whoops! Almost there
- Incorrect
- No, that was not the right answer

Appendix F

Data Analysis Results (written in Python)

OLS regression results for emotional response based on condition

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               L1_score   R-squared:                       0.163
Model:                            OLS   Adj. R-squared:                  0.148
Method:                 Least Squares   F-statistic:                     10.53
Date:                Sun, 23 Apr 2023   Prob (F-statistic):            0.00202
Time:                        14:16:24   Log-Likelihood:                -182.32
No. Observations:                  56   AIC:                             368.6
Df Residuals:                      54   BIC:                             372.7
Df Model:                           1
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                 16.5926      1.230     13.489      0.000      14.127      19.059
condition_positive     5.5453      1.709      3.244      0.002       2.118       8.972
==============================================================================
Omnibus:                        0.017   Durbin-Watson:                   2.054
Prob(Omnibus):                  0.991   Jarque-Bera (JB):                0.128
Skew:                          -0.039   Prob(JB):                        0.938
Kurtosis:                       2.780   Cond. No.                         2.66
==============================================================================
```

Independent t-test results for positive mistake message preference ratings based on math anxiety level (low vs high)

Variable	N	Mean	SD	SE	95% Conf.	Interval
Low	38.0	3.826316	1.622684	0.263234	3.292953	4.359679
High	19.0	4.824561	1.412154	0.323970	4.143925	5.505198
combined	57.0	4.159064	1.614346	0.213825	3.730721	4.587408

Independent t-test	results
Difference (Low - High) =	-0.9982
Degrees of freedom =	55.0000
t =	-2.2819
Two side test p value =	0.0264
Difference < 0 p value =	0.0132
Difference > 0 p value =	0.9868
Cohen's d =	-0.6412
Hedge's g =	-0.6324
Glass's delta1 =	-0.6152
Point-Biserial r =	-0.2941

Independent t-test results for neutral mistake message preference ratings based on math anxiety level (low vs high)

Variable	N	Mean	SD	SE	95% Conf.	Interval
Low	38.0	4.419298	1.098474	0.178196	4.058239	4.780358
High	19.0	3.294737	1.464530	0.335986	2.588856	4.000618
combined	57.0	4.044444	1.331432	0.176352	3.691168	4.397721

Independent t-test	results
Difference (Low - High) =	1.1246
Degrees of freedom =	55.0000
t =	3.2531
Two side test p value =	0.0020
Difference < 0 p value =	0.9990
Difference > 0 p value =	0.0010
Cohen's d =	0.9140
Hedge's g =	0.9015
Glass's delta1 =	1.0237
Point-Biserial r =	0.4017

Two-Way ANOVA results for math task score based on the condition tone and math anxiety level

	sum_sq	df	F	PR(>F)
C(AMAS_avg_level)	29.833533	1.0	4.915735	0.030925
C(condition)	3.421252	1.0	0.563727	0.456083
C(AMAS_avg_level):C(condition)	3.343451	1.0	0.550908	0.461224
Residual	321.656349	53.0	NaN	NaN