

Exploration of the Various Factors that Facilitate Learning on a
Computer-Based Learning Platform(CBLP)

by
Ashish Gurung

A Dissertation
Submitted to the Faculty
of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy
in
Computer Science

August 2023

APPROVED:

Prof. Neil T. Heffernan, Major Advisor, Department of CS, WPI

Prof. Joseph E. Beck, Dissertation Committee, Department of CS, WPI

Prof. Lane T. Harrision, Dissertation Committee, Department of CS, WPI

Prof. Adam C. Sales, Dissertation Committee, Department of MA, WPI

Prof. Anthony F. Botelho, Dissertation Committee, Department of EdTech, UF

ABSTRACT

The integration of technology into education via Computer Based Learning Platforms (CBLPs) has marked a significant shift towards a multitude of innovative strategies, specifically designed to enrich the teaching-learning experience. These platforms have transformed learning for students by harnessing the power of automation to facilitate real-time formative assessments, provide immediate support, and create a more immersive learning experience. For educators, CBLPs serve as a tool for optimizing both their time and resources by automating time-intensive tasks such as grading and report generation. This optimization frees up invaluable time and resources, which can then be redirected towards more constructive aspects of teaching. Moreover, these platforms generate insightful data on students' needs, aiding educators in tailoring their instruction more effectively, and consequently enhancing the efficacy of their pedagogical strategies. Despite these promising advances, a substantial amount of work remains to fully utilize the potential of technology in education, thereby setting the stage for fruitful future research and development.

The increasing adoption of Computer Based Learning Platforms (CBLPs) has led to a concomitant growth in the availability of data on learners and educators. This has enhanced our capacity to analyze and comprehend both learner and teacher behaviors. The abundance of this data has stimulated educational researchers to explore, often at scale, various strategies aimed at optimizing the learning process. These endeavors encompass the use of research methodologies in Learning Analytics, Educational Data Mining, Artificial Intelligence in Education, and Learning Experience/Human Computer Interaction. This surge of data-driven methodologies in the educational paradigm has led to unprecedented opportunities for exploration and understanding of effective teaching and learning strategies in the field of educational technology, ultimately fostering a greater potential for transformative change in the way we approach education.

This dissertation explores the features of ASSISTments, a widely used Computer-Based Learning Platform (CBLP), and examines their impact on enhancing learning outcomes. ASSISTments, primarily used by teachers in the United States for middle school mathematics instruction, exemplifies the integration of technology in education.

In Part I, this dissertation scrutinizes the effectiveness of various instructional interventions conducted at scale within the ASSISTments platform. This includes an analysis of feedback on commonly made mistakes, different types of motivational messages designed to encourage positive learning behaviors, and a comparison of fill-in-the-blank problems versus multiple-choice questions.

Part II extends the feedback discussion, exploring crowdsourcing as a promising approach. It outlines the process of involving curriculum experts to train and oversee the crowdsourced feedback process, aiming

to balance the risks and benefits of this approach, especially given the sensitivity of providing instructional feedback to young learners.

In Part III, the dissertation investigates the complexities and methods of automating grading and feedback generation for open-ended student responses, examining the issue from the perspective of considerate and fair grading practices. Furthermore, it explores teacher grading behavior, studying how students' demographic characteristics, such as gender and ethnicity, or their performance on past assignments can affect teachers' grading decisions.

In Part IV, the dissertation delves into the development of a classroom orchestration tool aimed at enhancing teaching. Initially, it explores the development of a low-fidelity detector, implemented post hoc, which analyzes student behavior in response to system-provided hints. The investigation then turns to the susceptibility of the Bayesian Knowledge Tracing algorithm to detector rot — a phenomenon where a model's performance degrades over time due to factors like covariate shift, overfitting, and systemic changes in the CBLP. Lastly, it reports on the creation of LIVE-CHART: Live Interactive Visual Environment for Creating Heightened Awareness and Responsiveness for Teachers, a prototype designed to provide real-time insights into student progress.

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to my family—my mother, father, Nani and Sundar kaka—for their relentless support and unyielding belief in me through this period of my life. I also wish to acknowledge my grandparents—Dan, Dil, Jas, and Aita—who raised me to be a good human, I hope I am doing a good job :). A special acknowledgement to my grandfather Dan, who instilled the value of education in me from an early age, and continues to be a guiding light in my pursuit of knowledge.

I would also like to extend my gratitude to the countless individuals who have made this dissertation possible through their invaluable support, guidance, and collaboration. My heartfelt thanks go to all those listed below, and to anyone whom I may have inadvertently overlooked but who nonetheless contributed to my journey, the **** is for you-thank you.

Neil	Anthony	Joe	Adam	Lane	Korinn	Ryan B.
Angela	Nicole	Erin	Stacy	Ji-Eun	Cindy	Joan
Sami	Rahul	Kirk	Andrew	Hannah	Haim	Ethan
March	Paul	Avery	John	Hilary	Anmesh	Anzana
Mikesh	Sobit	Bhakta	Jyoti	Prayan	Prishu	Joe St.
Aditya	Alphonsus	Anurag	Celso	Prajjwal	Prakash	Kala
Dilu	Jessica	Bhuwan	Parmita	David	Chris	Ryan E.
Pratik	Prajjwal Dai	Colleen	Becram	Ke	Lachhuman	Topraj
Ram	Hari	Janaki	Amrita	Rabi	Dovan	Anjana Di

FUNDING

The work presented in this dissertation has been funded by a number avenues including NSF (2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428), IES (R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, R305A120125 R305R220012), GAANN (P200A180088 P200A150306), EIR (U411B190024 S411B210024, S411B220024), ONR (N00014-18-1-2768), NHI (via a SBIR R44GM146483), Schmidt Futures, BMGF, CZI, Arnold, Hewlett and a \$180,000 anonymous donation. None of the opinions expressed here are those of the funders.

TABLE OF CONTENTS

	Page
LIST OF TABLES	xiii
LIST OF FIGURES	xvii
CHAPTER	
I Analyzing Instructional Interventions at Scale	1
1 IDENTIFICATION, EXPLORATION, AND REMEDIATION: CAN TEACHERS PREDICT COMMON WRONG ANSWERS?	2
1.1 Introduction	2
1.2 Related Works	4
1.3 Methodology	5
1.3.1 Study Design	6
1.3.2 Description of Dataset	7
1.4 Defining Common Wrong Answers	9
1.4.1 Identifying & Analyzing CWAs	10
1.4.2 Results of Identifying CWAs	12
1.5 Analysis of the effectiveness of CWAFs	12
1.5.1 Descriptive Statistics	13
1.5.2 Methods to Examine Effects of CWAF on Learning	14
1.5.3 Results on the Effectiveness of CWAFs	14
1.6 Exploring Personalization Effects	15
1.6.1 Identifying Heterogeneous Treatment Effects	15
1.6.2 Results Exploring Personalization	16
1.7 Discussion and Future Works	18
1.8 Conclusion	20
2 IMPACT OF NON-COGNITIVE INTERVENTIONS ON STUDENT LEARNING BEHAV- IORS AND OUTCOMES: AN ANALYSIS OF SEVEN LARGE-SCALE EXPERIMENTAL INTERVENTIONS	21
2.1 Introduction	21
2.2 Prior Work/Background	23
2.2.1 Mindset Theory	23

CHAPTER	Page
2.2.2	Achievement Emotions & Control-Value Theory 24
2.2.3	Social Comparison Theory & Self-Concept 25
2.2.4	Metacognition 26
2.3	Current Study 27
2.3.1	Research Questions 27
2.4	Method 28
2.4.1	Experimental Interventions 28
2.4.2	Data 30
2.4.3	Analytic Approach 30
2.5	Results 33
2.5.1	Embracing Mistakes Intervention 33
2.5.2	Inspirational Quotes Intervention 34
2.5.3	Social Comparison Intervention 35
2.5.4	Emotion Labeling Intervention 35
2.5.5	Confidence Judgments Intervention 36
2.6	Discussion 36
2.7	Conclusion 38
3	MULTIPLE CHOICE VS. FILL-IN PROBLEMS: THE TRADE-OFF BETWEEN SCALA- BILITY AND LEARNING 39
3.1	Introduction 39
3.2	Prior Works 42
3.2.1	Mastery-Based Learning Activities 43
3.3	Study Design 44
3.3.1	Description of Dataset 45
3.3.2	Attrition 46
3.3.3	Descriptive Statistics 47
3.4	Analysis 1: Impact of Fill-In Problems on Student Learning 47
3.4.1	Methods 48
3.4.2	Results 49

CHAPTER	Page
3.5 Analysis 2: Effect of Fill-In Problems During Atypical Learning Periods	50
3.5.1 Methods	51
3.5.2 Results	51
3.6 Analysis 3: Personalization Effect of Fill-In Problems	52
3.6.1 Methods	53
3.6.2 Results	53
3.7 Discussion and Future Works	55
3.8 Conclusion	57
II Crowdsourcing Instruction at Scale	59
4 HOW COMMON ARE COMMON WRONG ANSWERS? CROWDSOURCING REMEDIA- TION AT SCALE	60
4.1 Introduction	60
4.1.1 Research Questions	62
4.2 Background	62
4.2.1 Common Wrong Answers	62
4.2.2 Feedback Intervention	63
4.2.3 Common Wrong Answer Feedback	65
4.2.4 Crowdsourcing Instruction	65
4.3 Exploring Common Wrong Answers	66
4.4 Task Abstraction	68
4.4.1 Goal Analysis	69
4.4.2 Task Analysis	70
4.5 Crowdsourcing Common Wrong Answer Feedback	71
4.6 Implementing Common Wrong Answer Feedback	73
4.6.1 Experimental Design	73
4.6.2 Dataset	73
4.6.3 Evaluating the Effectiveness of Common Wrong Answer Feedback	74
4.7 Discussion and Future works	75
4.8 Conclusion	78

CHAPTER	Page
III Use of Automated Grading and Feedback Generation on Open Response Problems in Mathematics	80
5 INVESTIGATING PATTERNS OF TONE AND SENTIMENT IN TEACHER WRITTEN FEEDBACK MESSAGES	81
5.1 Introduction	81
5.2 Background	82
5.3 Dataset	84
5.4 Analysis 1: Sentiment Analysis in Mathematics	84
5.5 Analysis 2: Analyzing Tone using Punctuation Marks	87
5.5.1 Results of Analysis 2	88
5.6 Analysis 3: Comparing Sentiment and Tone	90
5.6.1 Results of Analysis 3	91
5.7 Conclusions and Future Work	92
6 AUTO-SCORING STUDENT RESPONSES WITH IMAGES IN MATHEMATICS	93
6.1 Introduction	93
6.2 Related Works	95
6.2.1 Automated Scoring Models	95
6.2.2 Methods for Image Analysis and Representation	96
6.2.3 The SBERT-Canberra Model	96
6.3 Dataset	97
6.4 Methodology	98
6.4.1 CLIP-Text Method	98
6.4.2 CLIP-Image Method	99
6.4.3 CLIP-OCR Method	99
6.5 Results	99
6.6 Error Analysis	101
6.6.1 Results of Error Analysis	102
6.7 Limitations and Future Works	102
6.8 Conclusion	104

7	CONSIDERATE, UNFAIR, OR JUST FATIGUED? EXAMINING FACTORS THAT IMPACT TEACHERS	105
7.1	Introduction	105
7.2	Background	107
7.2.1	Open-Ended Problems in ASSISTments	108
7.3	Study 1: Examining Grading Differences When the Student is Anonymized	109
7.4	Study 2: Exploring Related Factors of Student Assessment	112
7.4.1	Description of the Dataset	112
7.4.2	Factors Related to Student Grades	114
7.5	Study 3: Potential Impacts of Fatigue on Grading	116
7.6	Discussion	116
7.7	Limitations Future work	118
7.8	Conclusion	118
8	EXPLORING THE INFLUENCE OF ANONYMITY AND PRIOR-PERFORMANCE ON TEACHER GRADING BEHAVIOR.....	120
8.1	Introduction	121
8.2	Background	123
8.2.1	Halo Effect	123
8.2.2	Influence of Student Identity	124
8.2.3	Influence of Teacher Identity and Other Teacher Level Factors	126
8.2.4	Procedural Patterns in Teacher Grading Behavior	127
8.2.5	Open Ended Problems in Computer Based Learning Platforms	128
8.3	Methodology	129
8.3.1	Participants	130
8.3.2	Study Design	131
8.3.3	Description of Dataset	134
8.4	Intra-Rater Reliability	135
8.5	Main Effects of Anonymization and Prior Performance Information	137
8.5.1	Analysis Plan	137

CHAPTER	Page
8.5.2 Result	138
8.6 Influence of Gender and Ethnicity on Teacher Grades	139
8.6.1 Sub Effects	139
8.6.2 Sub Effects and Learner Gender	140
8.6.3 Sub Effects and Learner Ethnicity	140
8.6.4 Sub Effects, Learner Gender and Ethnicity	142
8.6.5 Results	143
8.7 Discussion	144
8.8 Conclusion	146
IV Facilitating Classroom Orchestration and Teaching Augmentation	148
9 EXAMINING STUDENT EFFORT ON HELP THROUGH RESPONSE TIME DECOMPO- SITION	149
9.1 Introduction	149
9.2 Background	151
9.3 Theoretical Framework behind decomposition of help usage	153
9.4 Description of DataSet	154
9.4.1 Action pairs considered	156
9.5 Exploratory Analyses	157
9.5.1 Analyzing action pairs	157
9.5.2 Examining Potential Systemic Causes	159
9.6 Examining Student Effort	162
9.6.1 Defining Effort	162
9.6.2 Modeling Student Effort	162
9.6.3 Exploring the Relationship Between Effort and Performance Metrics	163
9.7 Results	164
9.8 Discussion and future works	165
9.9 Conclusion	166
10 KNOWLEDGE TRACING OVER TIME: A LONGITUDINAL ANALYSIS	168
10.1 Introduction	168

CHAPTER	Page
10.1.1 COVID-19 Pandemic.....	170
10.2 Related Work.....	170
10.3 Methods	171
10.3.1 Data Collection	171
10.3.2 Student Modeling	172
10.4 Results	173
10.4.1 Robustness Over Time (RQ1).....	173
10.4.2 Complexity (RQ2)	174
10.4.3 Sudden Shifts: Pandemic Analysis (RQ3)	174
10.5 Discussion	175
11 LIVE-CHART: LIVE INTERACTIVE VISUAL ENVIRONMENT FOR CREATING HEIGHT- ENED AWARENESS AND RESPONSIVENESS FOR TEACHERS	177
11.1 Introduction	177
11.2 Theoretical Framework	179
11.3 Examining Existing Teaching Augmentation tools	182
11.4 Task Abstraction.....	183
11.4.1 Goal Analysis.....	184
11.4.2 Task Analysis	184
11.5 Implementation through LIVE-CHART	185
11.5.1 Data Characterization	186
11.6 Visualization and Interaction Design.....	189
11.6.1 Visualization Design	189
11.6.2 Interaction Design	192
11.7 Usability Study	193
11.7.1 Real Time usage	195
11.7.2 Takeaways.....	196
11.8 Future work and Open Questions	197
11.9 Conclusion	198

CHAPTER	Page
REFERENCES	199
APPENDIX	
A SUPPLEMENTARY MATERIALS FOR CHAPTER 8 “EXPLORING THE INFLUENCE OF ANONYMITY AND PRIOR-PERFORMANCE ON TEACHER GRADING BEHAVIOR”	222
A.1 Study Design	222
A.2 Additional Materials for Replication	222
A.2.1 Intra-rater reliability across condition per teacher	225
A.3 Main Effects	227
A.4 Sub Effects	229
A.4.1 Sub Effects and Learner Gender	230
A.4.2 Sub Effects and Learner Ethnicity	230
A.4.3 Sub Effects, Learner Gender and Ethnicity	235
B SUPPLEMENTARY MATERIALS FOR CHAPTER 11 “LIVE-CHART: LIVE INTERACTIVE VISUAL ENVIRONMENT FOR CREATING HEIGHTENED AWARENESS AND RESPONSIVENESS FOR TEACHERS ”	238
B.1 Task Abstraction	238
B.1.1 Goal Analysis	238
B.1.2 Task Analysis	238
B.2 Custom Seating Arrangement	238
B.3 Alternative Seating Arrangements	238

LIST OF TABLES

Table	Page
1.1 Filtered list of teachers, classes, assignments and student working on the two problem set.	9
1.2 CWAs identified by teachers by analyzing the problems.....	10
1.3 Analyzing CWAs that were made by the students with a threshold $N \geq 5$	11
1.4 Analyzing CWAs that were made by the students with a threshold $N \geq 10$	11
1.5 Descriptive Statistics of the experiment across the control and treatment condition for the two activites.	13
1.6 Effect of Common Wrong Answer Feedback (CWAF) on Mastery by Activity	15
1.7 Effect of Common Wrong Answer Feedback (CWAF) on Wheel-Spinning by Activity	16
1.8 Models Estimating Interactions Between Prior Performance and Common Wrong Answer Feedback (CWAF) Effects on Mastery by Activity	17
2.1 Intervention descriptions and theoretical basis	29
2.2 Experimental Sample Sizes	30
3.1 Student Level Attrition	47
3.2 Descriptive Statistics	47
3.3 Analysis 1 Results.....	50
3.4 Analysis 2 Results.....	52
3.5 Analysis 3 Results.....	54
4.1 Summary of Total Problems and Problems with CWAs. The problems with CWAs met our threshold of more than 20 students working on the problem in two or more academic years. ...	66
4.2 Common Wrong Answer by Student Count on the second problem as presented in figure 4.2. The threshold for the CWA requirement was met in 4 of the 5 academic years from ‘15-‘20. The threshold required more than 20 students to work on the problem in each academic year with more than 10 students making the same CWA.	68
4.3 Fundamental goals of a crowdsourcing tool.	70
4.4 Task analysis deconstructing the feature requirements of each sub-goal.	71
4.5 Exploring the effectiveness of CWAF by using next-problem-correctness(binary) as a de- pendent measure for the same set of Common Core Standards (within-skill) in consecutive problems.	76

4.6	Exploring the effectiveness of CWAF by using next-problem-correctness(binary) as a dependent measure within-assignment irrespective of the set of Common Core Standards associated with consecutive problems.	76
5.1	Most common mathematical words picked from a list of the top 100 most frequent words in the teacher feedback messages dataset categorized by their sentiment.	85
5.2	Some examples of Positive and Negative feedback messages from teachers, their sentiments with and without math terms, and their score.	86
5.3	Percent of feedback with commonly used punctuation marks across each of the score categories.	87
5.4	The resulting model coefficients for the multi-level logistic regression model on the use of question marks	88
5.5	The resulting model coefficients for the logistic regression model on the use of exclamation marks categorized according to low and high score categories.	89
5.6	The resulting model coefficients for the logistic regression model on the sentiment (positive = 1 and negative = 0) of feedback messages.	91
6.1	Model Performance compared to the auto-scoring methods developed in the prior works [30] .	100
6.2	The resulting model coefficients for the linear regression model of error for the auto-scoring method, conducted as a part of the error analysis similar to the prior method from Baral et. al [30].	101
7.1	Exploring the grading behavior of teachers when they had access to students' identity vs. when students were anonymized using Linear Weighted Cohen's Kappa.	111
7.2	Filtered Action Pairs of Graded and Ungraded Student Responses. This table shows the dataset after filtering for instances where a student either initiated and completed a problem, or resumed a previously incomplete problem and made a submission.	114
7.3	Linear Regression and Mixed-Effect model coefficients observing assessment score.	115
8.1	Description of the three types of teacher categories participating in the study.	130
8.2	Student pseudonyms used in the experiment with the index. The index represents the order in which the names were arranged in the list and maps to the fixed order of responses within each subsample.	131
8.3	Description of the total problems in the unique teachers, problems and responses per categories.	135

8.4	Description of the feedback length (words) and time required (seconds) to provide a score and feedback per score in each condition.	135
8.5	Comparison of Original and Anonymized Scores for intra-rater Reliability of Category 1 Teachers' Response Grades: Replicating Findings from Prior Work [144] on the First Batch of Responses	136
8.6	Comparison of teacher scores across conditions in the factorial experiment for all teachers.	136
8.7	Exploring the main effects of the two factors, i.e., the influence of student ethnic names (pseudonyms) prior performance information on teacher grading behavior.	138
8.8	Exploring the sub effects of the 2×2 factorial design examining the influence of student identity and prior performance information on teacher grading behavior.	140
9.1	filtered Action Pairs of students who asked for a hint	156
9.2	the mean(μ) and standard deviation(σ) for the high and low effort clusters using Gaussian Mixture Modelling	163
9.3	Logistic Regression analysis exploring the relationship between effort and next problem correctness while controlling for prior percent correct ($R^2 = 0.048$)	164
9.4	Logistic Regression analysis exploring the relationship between effort and wheel spinning while controlling for prior completion ($R^2 = 0.091$)	165
9.5	Logistic Regression analysis exploring the relationship between effort and assignment completion while controlling for prior completion ($R^2 = 0.104$)	165
10.1	Dataset Information	171
10.2	Feature List	172
10.3	BKT cross-year analysis	173
10.4	BKT+Forgets cross-year analysis	173
10.5	Cross-Pandemic Analysis	175
11.1	Fundamental goals of a TA tool.....	185
11.2	Fundamental goals of a TA tool.....	186
11.3	A snippet of the problem information inside the data structure for assignment data.	188
11.4	A snippet of the student data where the student answered the problem correctly after asking for a hint from the system.	189

A.1	Comparison of Original and Anonymized Scores for intra-rater Reliability of Category 1 Teachers' Response Grades: Replicating Findings from Prior Work [144] on All Responses ...	224
A.2	Examining some of the grades and feedback of teacher 3 with relatively low intra-rater reliability.	224
A.3	Regular and Relaxed (Off by 1) intra-rater reliability across condition per teacher. The values ≤ 0.5 intra-rater reliability have been marked in red.	226
A.4	Comparing the estimation of main effects when using dummy coding compared with when using effect coding.....	228
A.5	Comparing sub-effects across genders using anonymized as the baseline to compare the distribution of grades post randomization.	230
A.6	Comparing sub-effects across conditions for different genders separately.	231
A.7	Comparing sub-effects across ethnicities using anonymized as the baseline to compare the distribution of grades post randomization.	232
A.8	Comparing sub-effects across ethnicities using anonymized as the baseline to compare the distribution of grades post randomization.	233
A.9	Comparing sub-effects across conditions for different ethnicities separately.	234
A.10	Comparing sub-effects across ethnicities for both genders using anonymized as the baseline to compare the distribution of grades post randomization.....	235
A.11	Comparing sub-effects across conditions for different ethnicities separately for boys.	236
A.12	Comparing sub-effects across conditions for different ethnicities separately for girls.	237
B.1	Fundamental goals of a TA tool.....	239
B.2	Fundamental goals of a TA tool.....	240

LIST OF FIGURES

Figure	Page
1.1 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.	8
1.2 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.....	8
1.3 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.	9
1.4 Comparing the effect of Common Wrong Answer Feedback (CWAFs) on mastery and wheel spinning behavior of students.....	15
1.5 Interaction between students’ prior percent correct and the predicted probability of mastery for the “Order of Operations activity” by condition.	18
2.1 This plot provides the effect size, standard deviation, and statistical significance for each of the experimental conditions on each outcome. Statistical significance is based on p-values adjusted using the Bonferroni-Holm procedure within each experiment.	34
3.1 Breakdown of the experimental design. It has two conditions where students are assigned a mastery based assignment with MCQs or Fill-In problems. Upon mastering the content students are asked to answer two Fill-In problems that with higher difficulty than the problems in the experiment.	45
3.2 Example Problems from “Problem Set 1: Greatest Common Factor”	46
3.3 Interaction effect of problem type and prior performance on likelihood of post test correctness.	55
4.1 A model of feedback for enhanced learning, taken from Hattie et al. (2007) [153].....	64
4.2 An example of two consecutive problems from ENY Grade 7 Module 3 Lesson 1 where both problems have the same set of Common Core Standards.	67
4.3 Teacher perspective, visualization of a problem from Illustrative Math curricula with Common Core standard 7.SP.C.8.b where a teacher has written feedback and a peer/moderator has reviewed it as well.	72

5.1	Plot showing the total number of feedback messages per score category grouped by the sentiment of feedback messages (a) and the average sentiment estimate across the first, second, third, and fourth sentences in the feedback message using continuous-valued outcomes from the model (b).	85
6.1	Simplified representation of the SBERT-Canberra method to generate a predicted score by identifying the most similar historic response to a given new student answer using Canberra distance within an embedding space.	94
6.2	Examples of image-based responses from students given in response to Open-ended math problems	98
7.1	The different types of open-ended problems on ASSISTments. (a) the main problem is open-ended (b) a multi-part problem where the second problem is open-ended.	109
7.2	The interface for teachers to grade students' responses on open-ended problems.	109
7.3	The interface for teachers to grade students' responses on open-ended problems.	117
8.1	Examples of open-ended problems in ASSISTments.	128
8.2	A screenshot from the grading tool where a teacher is grading a response from an African American female student Brianna Booker (pseudonym). We display the problem body, prior sub-parts, student's response and average performance on prior 5 assignments.	132
8.3	A visual representation of the 2×2 factorial randomized control trial.	134
8.4	A comparison of the difference in average scores between genders as teachers grade students in one of the four conditions in the 2×2 experimental design.	141
8.5	A comparison of the difference in average scores across ethnicities as teachers grade students in one of the four conditions in the 2×2 experimental design.	142
8.6	A comparison of the difference in average scores across student gender and ethnicities as teachers grade students in one of the four conditions in the 2×2 experimental design.	143
9.1	Visual representation of the student behaviour for a user interacting with a Computer-based learning platform.	153
9.2	distribution curve of (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs using natural log-transformed values of time taken for each action pair	158

9.3	distribution curve of (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs, when their first action was asking for hint after reading the problem, using natural log-transformed values of time taken for each action pair	159
9.4	There are too few instances of video hint for us to draw a conclusion but the data does seem to indicate that the type of hint does not influence the action pair response time	160
9.5	There are too few instances of video hint for us to draw a conclusion but the data does seem to indicate that the type of hint does not influence the action pair response time	161
9.6	The amount of time a user spends after getting a hint is the same for students who made a correct or an incorrect attempt	161
10.1	Means and 95% CIs for models trained and evaluated on the same year	174
10.2	Means and 95% CIs for models trained on one side of the pandemic and trained on the other ..	175
11.1	A visual representation of how the introduction of computers in the classroom has disrupted the learning experience in traditional classrooms.	180
11.2	A visual representation of how the introduction of computers in the classroom has disrupted the learning experience in traditional classrooms. A. Lumilo, B. SEAT, C. Fireflies, and D. MTDashboard	183
11.3	Class view: visualization of students as they work on their classwork in real-time. (a)priority dashboard that highlights students doing well or requiring attention in class, (b) students visualized according to their classroom seating arrangement, (c) The controls: teachers can manipulate the visualization timeline as the data is temporal, and (d) individual student view representing high level information for teacher to infer progress.	191
11.4	Student Detail View: visualization of individual student's work with detailed per-problem information in real-time. (a)dashboard indicating if the student currently has been flagged for requiring attention or doing well, (b) problem level information representing student performance in prior problems, (c) The controls: teachers can manipulate the visualization timeline as the data is temporal (d) detailed breakdown of the problem and the actions a student took while working on the problem.	192
11.5	Seating Arrangement: teachers can arrange the students in the class to reflect the seating arrangement of the class.	193

11.6	The class sizes of the teachers in the study.	194
11.7	Teacher preference for student arrangement during virtual and in-person classes.	194
A.1	A screenshot from the grading tool where a teacher is grading a response where the student is anonymized and prior performance information is not provided. We display the problem body, prior sub-parts, and student's response.	222
A.2	A screenshot from the grading tool where a teacher is grading a response where the student identity (pseudonym) is provided but prior performance information is hidden. We display the problem body, prior sub-parts, and student's response.	223
A.3	A screenshot from the grading tool where a teacher is grading a response where students are anonymized but we provide their prior performance. We display the problem body, prior sub-parts, student's response and average performance on prior 5 assignments.	223
A.4	A screenshot from the grading tool where a teacher is grading a response where the student identity (pseudonym) and prior performance information is provided. We display the problem body, prior sub-parts, student's response and average performance on prior 5 assignments.	224
A.5	A screenshot of the Google search results for the term "Equivalent Ratios."	225
A.6	A visual representation of the 2 × 2 factorial randomized control trial.	225
B.1	Seating Arrangement: teachers can arrange the students in the class to reflect the seating arrangement of the class.	239
B.2	Visualization of students in Alphabetical order.	241
B.3	Visualization of students in anonymized Alphabetical order.	241
B.4	Visualization of students in per-problem view.	242
B.5	Visualization of students in anonymized per-problem view.	242

Part I

Analyzing Instructional Interventions at Scale

Chapter 1

IDENTIFICATION, EXPLORATION, AND REMEDIATION: CAN TEACHERS PREDICT COMMON WRONG ANSWERS?

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.

Proper citation for this chapter is as follows:

Gurung, A., Baral, S., Vanacore, K.P., McReynolds, A.A., Kreisberg, H., Botelho, A.F., Shaw, S.T., & Heffernan, N.T. (2023). Identification, Exploration, and Remediation: Can Teachers Predict Common Wrong Answers? In LAK23: The 13th International Learning Analytics and Knowledge Conference (LAK 2023).

1.1 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” [22] 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 [160]. Within this, many teachers are able to use experiential knowledge to recognize the types of mistakes, sometimes referred to as “bugs” [53], 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., [53]). Brown and colleagues ([53]) 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 [257]. 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 [53, 359]. 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?

1.2 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 [54, 53, 394, 58, 387, 258]. Others have also focused on rectifying these errors through instruction [81, 329]. As such, Brown and colleagues [53] 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 [54], they ex-

plain 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 [341] 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. [329] 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 [249, 250]. A study from Vanlehn and colleagues [362] 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 [326]; this suggests that deeply ingrained misconceptions may be more difficult to rectify over time. Other works [198, 139, 319] 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. [319], report on the approach of error analysis leading to better knowledge retention over the traditional methods of learning mathematics.

1.3 Methodology

For all of our analyses, we utilize data that was collected from a randomized controlled trial designed to measure the learning impacts of CWA. 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.

1.3.1 Study Design

Exploring the effectiveness of CWAFs uses two activities on the ASSISTmetns platform[154]; 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 [79] 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. [180] 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. [290] 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 [18] 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 9th or 10th 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 [35].

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 1.1. 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 1.2, 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 1.3, 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”[362].

1.3.2 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 [18]. During this period, 587 middle school teachers in the United States assigned one or both mastery-based activities to 1283 of their classes

¹The templates for the CWAFs are publicly available at <https://osf.io/gjst9/>

²The 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
<p>Template 1:</p> <p>What is the solution to the expression below? $\%v\{a\} + \%v\{b\} \times \%v\{c\}$</p> <p>Example: What is the solution to the expression below? $7 + 4 \times 3$</p>	<p>Template 1:</p> <p>Solve for $\%v\{a\}$ $\%v\{c1\}\%v\{a\} + \%v\{c2\} = \%v\{c3\}$</p> <p>Example: Solve for a. $9a + 10 = 28$</p>
<p>Template 2:</p> <p>What is the solution to the expression below? $\%v\{a\} - \%v\{b\} \times \%v\{c\}$</p> <p>Example: What is the solution to the expression below? $5 - 2 \times 5$</p>	<p>Template 2:</p> <p>Solve for $\%v\{a\}$</p> <div style="border: 1px dashed gray; padding: 5px; display: inline-block;"> $\%v\{a\} + \%v\{c\} = \%v\{d\}$ </div> <p>Example: Solve for y. $\frac{y}{2} + 6 = 4$</p>

Figure 1.1: 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 1.2: 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.

resulting in 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

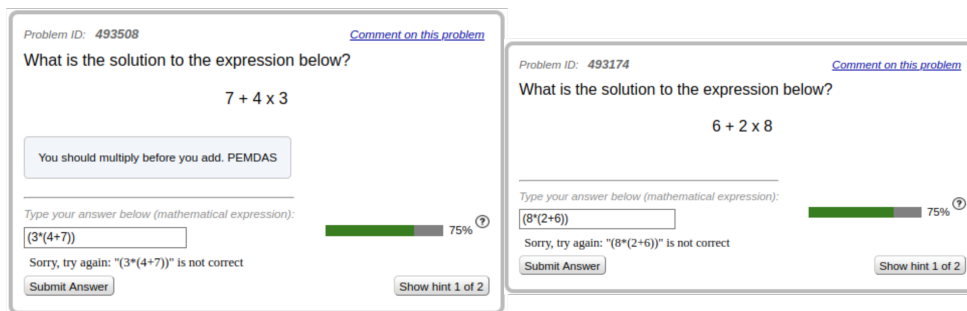


Figure 1.3: Example problems in treatment (problem on the left) and control (problem on the right) condition for “Order of Operations” activity. The CWAf is provided to students when they provide a CWA in the treatment condition.

student if their students initially performed poorly on the assignment. For instances where students worked on the mastery-based activity more than once, 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 1.1 lists the number of teachers, classes, assignments, and students after implementing the filtration procedures on the data.

Table 1.1: 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

1.4 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 1.2 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 1.2 & figure 1.3.

Table 1.2: CWAs identified by teachers by analyzing the problems.

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

1.4.1 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 1.3 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 1.3: Analyzing CWAs that were made by the students with a threshold $N \geq 5$.

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

Table 1.4 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 correctly identified. In contrast, teachers were unable to identify 54 (~21%) CWAs and provide appropriate CWAFs.

Table 1.4: Analyzing CWAs that were made by the students with a threshold $N \geq 10$.

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

1.4.2 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 : $\sim 32\%$ 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”, $\sim 63\%$ 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 : $\sim 21\%$ 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”, $\sim 39\%$ 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.

1.5 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 [35]; 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.

1.5.1 Descriptive Statistics

We evaluated the student data on the mastery-based activities and compared the problems to mastery, hint usage, average problem 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 1.5 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.

Table 1.5: Descriptive Statistics of the experiment across the control and treatment condition for the two activities.

	2-Step Equation				Order of Operations			
	Control		Treatment		Control		Treatment	
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>
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

1.5.2 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 more concise 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 [119].

The logit regressions were estimated because mastery and wheel spinning are binary outcomes. Equation 8.1 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 8.1 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} .

$$\text{logit}(\text{Outcome is True for Student } i \text{ with teacher } t) = \gamma_{00} + \gamma_{10}CWAF_i + \mu_{0t} \quad (1.1)$$

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 1.4 presents the treatment effects for each activity and learning outcome. In table 1.6 & table 1.7 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 ($\gamma_1 = -1.30$, SE = 0.06, $p = 0.027$) and a positive effect on the likelihood that students would wheel-spin during the activity ($\gamma_1 = 0.51$, 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 ($\gamma_1 = -0.51$, SE = 0.09, $p = 0.001$) and wheel-spinning ($\gamma_1 = 0.21$, SE = 0.06, $p < 0.001$). Yet, for the “Order of Operations”

activity, neither of the effects on mastery ($\gamma_1 = 0.22$, $SE = 0.14$, $p = 0.144$) nor wheel-spinning ($\gamma_1 = 0.30$, $SE = 0.27$, $p = 0.264$) were significant. Notably, the point estimate for the CWF 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.

Table 1.6: Effect of Common Wrong Answer Feedback (CWF) on Mastery by Activity

<i>Predictors</i>	Both Activities		2-Step Equations		Order of Operations	
	<i>Log-Odds</i>	<i>SE</i>	<i>Log-Odds</i>	<i>SE</i>	<i>Log-Odds</i>	<i>SE</i>
Intercept	2.95***	0.08	2.68***	0.09	3.66***	0.17
CWAFs (Treatment)	-0.13**	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	

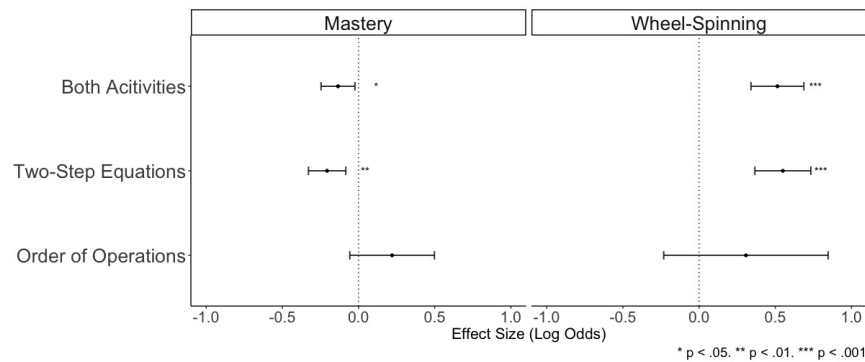


Figure 1.4: Comparing the effect of Common Wrong Answer Feedback (CWAFs) on mastery and wheel spinning behavior of students.

1.6 Exploring Personalization Effects

1.6.1 Identifying Heterogeneous Treatment Effects

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

Table 1.7: Effect of Common Wrong Answer Feedback (CWAF) on Wheel-Spinning by Activity

<i>Predictors</i>	Both Activities		2-Step Equations		Order of Operations	
	<i>Log-Odds</i>	<i>SE</i>	<i>Log-Odds</i>	<i>SE</i>	<i>Log-Odds</i>	<i>SE</i>
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	

CALP platform and the CWAF condition to the base model used in Section 1.5 (Equation 8.1). 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 from 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%.

1.6.2 Results Exploring Personalization

Overall there was a significant interaction between students' prior percent correct and the CWAFs condition. Table 1.8 displays the results for these models. For students with the mean prior accuracy, the effect of CWAF was negative ($\gamma_1 = -0.17$, $SE = 0.07$, $p = 0.017$). The interaction effect was also negative ($\gamma_2 = -0.11$,

Table 1.8: Models Estimating Interactions Between Prior Performance and Common Wrong Answer Feedback (CWAF) Effects on Mastery by Activity

<i>Predictors</i>	Both Activities		2-Step Equations		Order of Operations	
	<i>Log-Odds</i>	<i>SE</i>	<i>Log-Odds</i>	<i>SE</i>	<i>Log-Odds</i>	<i>SE</i>
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	

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 ($\gamma_1 = -0.18$, SE = 0.08, $p = 0.019$), but the interaction was not significant ($\gamma_2 = -0.08$, 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 ($\gamma_1 = -0.06$, SE = 0.18, $p = 0.756$), but the interaction was significant ($\gamma_2 = -0.36$, SE = 0.14, $p = 0.013$). Figure 3.3 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 1.5), 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.

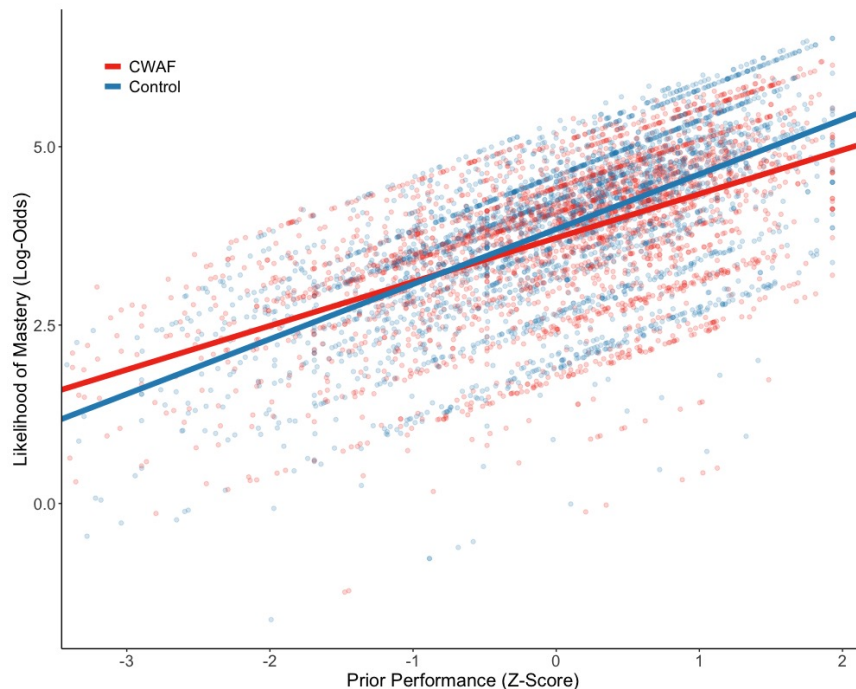


Figure 1.5: Interaction between students' prior percent correct and the predicted probability of mastery for the "Order of Operations activity" by condition.

1.7 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. [53] observed that students working on basic arithmetic problems can reach the same incorrect answer using different approaches, which required the tutors to 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 [142] 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 [362] 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 [53] 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.

1.8 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.

IMPACT OF NON-COGNITIVE INTERVENTIONS ON STUDENT LEARNING BEHAVIORS AND OUTCOMES: AN ANALYSIS OF SEVEN LARGE-SCALE EXPERIMENTAL INTERVENTIONS

As evidence grows supporting the importance of non-cognitive factors in learning, computer-assisted learning platforms increasingly incorporate non-academic interventions to influence student learning and learning related-behaviors. Non-cognitive interventions often attempt to influence students' mindset, motivation, or metacognitive reflection to impact learning behaviors and outcomes. In the current paper, we analyze data from five experiments, involving seven treatment conditions embedded in mastery-based learning activities hosted on a computer-assisted learning platform focused on middle school mathematics. Each treatment condition embodied a specific non-cognitive theoretical perspective. Over seven school years, 20,472 students participated in the experiments. We estimated the effects of each treatment condition on students' response time, hint usage, likelihood of mastering knowledge components, learning efficiency, and post-tests performance. Our analyses reveal a mix of both positive and negative treatment effects on student learning behaviors and performance. Few interventions impacted learning as assessed by the post-tests. These findings highlight the difficulty in positively influencing student learning behaviors and outcomes using non-cognitive interventions.

Proper citation for this chapter is as follows:

Vanacore, K.P., Gurung, A., McReynolds, A.A., Liu, A., Shaw, S.T., & Heffernan, N.T. (2023). Impact of Non-Cognitive Interventions on Student Learning Behaviors and Outcomes: An analysis of seven large-scale experimental inventions. In LAK23: The 13th International Learning Analytics and Knowledge Conference (LAK 2023).

2.1 Introduction

In recent decades, the use of computer-assisted learning platforms (CALPs) as a complement to traditional classroom instruction has increased dramatically [278, 173, 136]. Learning-experience designers often embed both academic and non-academic supports into CALPs to improve students' learning outcomes. Much of the causal research in learning analytics focuses on the effects of academic interventions in CALPs – such as hints, scaffolding, and performance based-feedback [288, 291, 273, 210] – but less attention has been

given to non-cognitive interventions that focus on students' mindsets, emotions, and motivation as they engage in learning activities. There is growing appreciation that "non-cognitive" aspects of education, such as students' motivation, persistence, and meta-cognition, are essential parts of the learning process [349, 183]. Non-cognitive factors are broadly defined as a set of skills and traits that are not analytic or intellectual but are academically relevant [315]. Although these factors are predictive of educational outcomes, attempts to implement non-cognitive interventions in various educational settings, from traditional classrooms and massive open online courses, have shown mixed results in influencing students' behaviors and learning outcomes [393, 146, 190]. As learning-experience designers incorporate non-cognitive interventions into CALPs, research on how these interventions impact students' learning behavior and outcomes is essential to optimizing digital educational environments.

One way to influence non-cognitive factors is by embedding small interventions within a learning activity. In non-educational fields, behavioral scientists have found that embedding simple interventions, such as messages or short activities, "nudge" people toward specific desirable behaviors [353, 233]. Similarly, education researchers studying the impact of 'nudges' meant to influence students' mindsets and motivation and thus impact their academic behaviors and performance have found mixed results [190, 238, 283]. In a review of social-psychological interventions in education [393], Yeager and Walton note that non-cognitive interventions can seem "magical" because of their potentially long-lasting impacts, but they warn against this view. Instead, they state that we must understand these interventions as "powerful but context-dependent tools" for them to be effective at scale. In that light, we seek to understand the potential impact of these tools in the context of CALPs.

Presently, we explore the use of small non-cognitive interventions in ASSISTments focused on middle school mathematics. Specifically, we investigate the impacts of five experimental interventions that use motivational messaging and small metacognitive exercises based on non-cognitive theories – growth mindset, achievement emotions and control-value, social comparison and self-concept, and metacognition – that have implications for learning. We examine how these interventions impact learning by looking at multiple dimensions of students learning processes. Our findings provide evidence that different types of non-cognitive interventions embedded in online learning systems affect students' learning behaviors and outcomes differently. Practically, these findings can also inform the design of non-cognitive supports within online learning systems that can meaningfully impact student engagement and learning.

2.2 Prior Work/Background

A growing number of studies have investigated the relations between non-cognitive factors and learning [204, 349, 315]. The degree to which these factors are associated with learning outcomes vary. For example, affective factors such as depression, feelings of well-being, and generalized anxiety generally show small correlations with academic performance, whereas domain-specific self-efficacy and domain-specific anxiety are highly correlated with academic achievement [349]. Non-cognitive interventions seek to impact learning by influencing non-cognitive factors, which have theoretical downstream effects on their learning behavior and performance. Below, we describe four common educational theories focused on non-cognitive factors and how they purportedly impact learning behaviors and achievement.

2.2.1 Mindset Theory

Mindset theory focuses on individuals' beliefs about how abilities can be developed [106]. Individuals can hold either "growth" mindsets, in which they believe their talents can be developed through effort, or "fixed" mindsets, in which they believe their talents are innate and inflexible. According to this theory, people with higher growth mindsets should be more adaptive, particularly when facing difficulties, leading to increased academic achievement.

Studies have found that mindset can influence how students view and engage in learning activities. One early study on mindset found that the mindset portrayed by adults when they praise students' motivation and performance on learning tasks [245]. Research suggests students' mindsets are implicated in academic performance, how they address changes, and respond to academic stress [390]. Growth mindsets may also have indirect effects on achievement through motivational factors, such as increasing motivation to learn [56, 305] and grit [272].

However, there has been considerable debate around the generalizability of growth mindset interventions, as studies have found that effects are highly variable depending on individuals and contexts [391]. Sisk and colleagues [340] conducted two meta-analyses: one on the correlation between mindset and academic achievement (273 studies, $n = 365,915$) and one on the effect of mindset interventions on academic achievement (43 studies, $n = 57,155$). They found that over half of the effect sizes reviewed in their meta-analyses were not significant, and those that were significant had an average weak effect along with high degrees of heterogeneity. Specifically, students from lower socioeconomic backgrounds or those who were academically at-risk generally benefited, which has also been reflected in other large-scale studies. For example, a growth mindset predicted achievement in a national sample of 10th grade students in Chile and helped to

offset the negative effects of poverty on achievement [72]. Similarly, a large-scale randomized controlled trial investigating the effect of a short online mindset intervention on lower-achieving 9th grade students also found positive impacts on students' math GPA and enrollment in advanced math classes [392]. Still, other recent large-scale studies suggest that growth mindset may only be predictive of achievement among wealthy students and not those from less advantaged families [189]. Thus, it is still unclear how different implementations of growth mindsets will generalize when implemented into a large-scale CALP.

A key aspect of growth mindset is how students respond to failure during learning. Students with growth mindsets should view mistakes as learning opportunities, whereas students with fixed mindsets may view the same mistake as an indication of their poor ability. A study investigating undergraduates in a STEM course found that failure perception was a significant factor associated with changes in students' mindsets over the semester [211]. Students with fixed mindsets at the start of the semester were likely to report academic struggles and shift away further from growth mindsets throughout the semester. Alternatively, students who started the semester with growth mindsets and continued to hold higher growth mindsets throughout the semester reported lower levels of academic struggle during the semester. This suggests that how students perceive difficulty in learning influences whether they struggle academically.

In order to help students access potential benefits of growth mindsets, many interventions provide scripts for teachers that re-frame students' mistakes as opportunities to learn [324]. In educational games, growth mindset interventions have also increased overall gameplay, positive strategies, and persistence after challenges [260]. As such, growth mindset interventions focused on changing students' perceptions of failure while they engage with CALPs may be effective for improving learning behaviors and outcomes.

2.2.2 Achievement Emotions & Control-Value Theory

Achievement emotions are those tied to achievement activities or outcomes [276]. The specific emotions that individuals feel during or as a result of learning activities can have significant effects on learning and performance [277, 295]. As part of his Control-Value theory, Pekrun *et al.* [277] posited that individuals' achievement emotions are proximally determined by their appraisals of control (*i.e.*, their perceived influence over actions and outcomes) and value (*i.e.*, their perceived importance of success). The theory also classified 17 achievement emotions based on their valence (positive vs. negative), activation (activating vs. deactivating), and object focus (activity vs. outcome). For example, joy is classified as a positive, activating emotion focused on outcomes, and frustration is classified as a negative, deactivating emotion focused on activities.

Different emotions have also been found to relate differently with student achievement [62] through mech-

anisms such as consuming cognitive resources [112, 230], promoting certain strategies [73], or supporting interest and motivation to perform tasks. In general, an emotion's impact on learning depends on where it falls within Pekrun's taxonomy. Positive activating emotions are positively correlated with learning and performance, including outcomes of interest, effort invested, self-regulation of learning, grades, and test scores [212, 277]. Meanwhile, negative deactivating emotions can reduce cognitive resources available for tasks or lead to superficial information processing [276]. However, not all negative emotions are necessarily bad for learning outcomes; for example, a negative-activating emotion such as confusion can benefit learning outcomes if students overcome their confusion [108].

Much prior research on achievement emotions has taken place in controlled lab settings; therefore, it is unclear how non-cognitive interventions focused on achievement emotions will generalize in classrooms and at scale. A potential target for non-cognitive interventions is to promote positive, activating emotions such as joy or hope, which are associated with positive learning outcomes. Another option is to teach students to become more aware of their emotions during or after learning activities to help them identify and regulate those emotions. This emotion labeling is considered part of students' emotion knowledge (*i.e.*, the ability to perceive and label emotions in oneself accurately and others), which is considered a critical non-cognitive skill for students to learn. For example, a meta-analysis of 49 studies with students from ages 3-12 found that emotion knowledge of other people's emotions was correlated with academic performance with an effect size of .32, with stronger associations among middle-class children [367]. However, few studies have investigated how promoting emotion-labeling behaviors can impact their learning outcomes and behaviors.

2.2.3 *Social Comparison Theory & Self-Concept*

Social comparison describes any processes through which individuals relate their abilities to others [96]. People engage in social comparison for various reasons, including self-evaluation, self-improvement (to improve their skills), or self-enhancement (to protect or improve their self-esteem). In general, people compare themselves with people who are superior to them in some way, with such comparisons resulting in worsened moods and lower ability appraisals. In contrast, comparisons with people who are lower in ability result in more positive outcomes but are rarer [129]. This tendency has been leveraged by behavioral scientists to influence various behaviors, such as reducing energy conception [6, 248], health-related decisions [218, 233], and tax compliance [13]. These interventions utilize normative data on groups, to nudge individuals towards or away from certain behaviors.

In terms of academic achievement, students' academic self-concepts – perceptions of one's academic abil-

ities – are partly determined by social comparisons of their achievements against their peers [222]. In turn, academic self-concepts have been related to academic achievement [310, 239]. In a classroom context, students can explicitly compare their achievements against individual students while also implicitly measuring themselves against the perceived average ability and performance of their entire school, grade, or classroom [342]. Ability grouping or tracking within schools can further complicate which comparison groups are most salient for students. Indeed, students' social comparisons at the school- or classroom-level have significant effects on students' self-concept. Average school or classroom ability can affect students' individual student's self-concepts even when individual achievement is controlled [311, 223]. Thus, interventions that manipulate the target of social comparison and students' standing relative to a comparison group may be one way to bolster students' academic self-concepts and improve subsequent learning outcomes.

2.2.4 *Metacognition*

Metacognition encompasses individuals' knowledge and regulation of their cognition [121, 363]. In terms of academic achievement, having higher metacognitive knowledge can help students understand the factors that affect their academic outcomes and to plan, monitor, and evaluate their learning. Supporting metacognitive strategies during learning activities (*e.g.*, reflecting on one's knowledge prior to or after a learning activity) can help students to think about their learning process [55]. Several prior studies have shown that greater use of metacognitive strategies is positively and strongly associated with higher academic achievement [60, 259, 355, 368]. For example, in the 2009 PISA dataset of 15-year-old students across 65 countries, metacognitive strategies significantly predicted academic achievement when controlling for SES and gender [60].

Confidence judgments are one metacognitive strategy that can help students assess their state of knowledge [231]. Judgments conducted before learning reflect students' perceptions of their current knowledge and ease of learning, which are important components of students' metacognitive self-monitoring. Such judgments require students to think about the requisite knowledge for the task, their own abilities, and the steps they need to take for requisite knowledge and their abilities to align. Teaching students to self-monitor and make knowledge judgments can increase their overall test confidence; however, there is a risk that it can also make them overconfident in their abilities [172]. Thus, non-cognitive interventions that encourage students to judge their confidence while providing timely feedback during problem-solving may be another method for improving their metacognition, which should subsequently ease their learning processes and improve performance.

2.3 Current Study

In the current study, we evaluate the impact of non-cognitive interventions embedded into mastery-learning activities on student learning and learning-related behaviors. The study includes five experiments conducted through ASSISTments. This ASSISTments platform allows researchers to embed experiments into mastery-based learning activities and problem sets. To date, over 80 experiments have been run through the system. To determine which experiments to include in our analyses, we reviewed all experiments ($n = 14$) which did not include manipulations with academic features (hints, feedback, problem type, *etc.*). Of these experiments, five were included in the current study because they were conducted with a non-cognitive theoretical basis (*e.g.*, Growth Mindset, social comparison, *etc.*) and included a learning outcome to be used as one of the dependent measures. Each of the selected experiments were conducted in a mastery-based learning activity focused on a specific knowledge component, in which students completed problems until they mastered that knowledge component [344]. This selection process was conducted before extracting the data from the ASSISTments database. Prior to analyzing the data, we preregistered our analyses through the Open Science Foundation (OSF Link). Data and code for the analyse can be found on GitHub (GitHub Link).

2.3.1 Research Questions

As we are interested in understanding whether non-cognitive interventions affect learning-related behaviors as well as learning measures as outcomes, our research questions focus on a variety of different variables as outcomes:

RQ 1 Did each intervention increase students' initial response time when solving a problem?

As response time (defined as the time between viewing the problem and either submitting a response or requesting a hint) is positively correlated with performance [66, 143, 195], we analyzed whether each intervention impacted the amount of time between viewing the problem and taking an action (response time) as a learning-related outcome. This allows us to evaluate whether the interventions cause students to slow down and consider the problem prior to acting, an action that is related to learning and performance [202].

RQ 2 Did each intervention increase the likelihood of students engaging in hint usage during the activity?

Providing access to hints has a positive effect on student performance [288, 273], so the likelihood

that students utilized hints as a help-seeking behavior was included as an outcome. This allows us to measure the extent to which the interventions cause students to seek assistance as they worked through the activity.

RQ 3 Did each intervention increase the likelihood that students completed the activity by mastering the knowledge component?

The goal of each activity in ASSISTments is to master the knowledge components, so we evaluate each intervention's impact on the likelihood that the students reach this goal. Notably, mastery can be viewed as a product of student knowledge entering the activity, learning during the activity, and their willingness to persist through the activity.

RQ 4 Did each intervention impact student learning as measured by their efficiency in mastering the knowledge component of the activity and performance on a post-test?

Students required different numbers of problems before reaching mastery and may have experienced different levels of learning during the activity. Therefore, we examined whether the interventions impacted the efficiency in which students learned by examining the difference in the number of problems to mastery, and we evaluated how much learning they experienced during the activity by assessing differences in their performance on a post-test.

2.4 Method

2.4.1 *Experimental Interventions*

The five experimental interventions were conducted over eight problem sets. Two of the experiments had multiple treatment conditions, producing a total of seven experimental treatments. None of the interventions overlapped; only one experiment was conducted in each problem set, so students could only be in one intervention at a time. All experiments included a business-as-usual control condition, in which the students worked on the mastery-based learning activities and problem sets without any non-cognitive intervention.

The activities focused on different mathematical skills commonly covered as part of the United States middle school curricula (grades 7-8). These skills include adding decimals, adding and subtracting fractions, percentages, geometry, probability, permutations, and combinations.

Table 2.1 ¹ provides descriptions of each experiment, a label of their theoretical basis, indications of

¹Indicates whether the intervention was administered before the mastery-based learning activity, or while the students were completing problems in the activity.

Table 2.1: Intervention descriptions and theoretical basis

Treatment Conditions	Theory	Description	Administration		Number of Problem Sets
			Before	During	
Embracing Mistakes					
<i>Image</i>	Growth Mindset	Students were exposed to an image that said “Keep Calm and Learn From Your Mistakes” and a written message that encouraged students to reappraise their mistakes as opportunities to learn prior to starting the mastery learning activity	Yes	No	2
	Growth Mindset	Students were exposed to a video that encourages students to reappraise their mistakes as opportunities to learn prior to starting the mastery learning activity	Yes	No	2
<i>Video</i>					
Inspirational Quotes	Achievement Emotions & Control-Value Theory	Students were provided with positive messages and inspirational quotes from famous people after they submitted answers during the mastery learning activity	No	Yes	2
Social Comparison					
<i>Performance</i>	Social Comparison	Students were told average number of problems completed by their peers to master the content prior to starting the mastery learning activity	Yes	No	1
	Social Comparison	Students were told the percentage of their peers who used hints in the problem set prior to starting the mastery learning activity	Yes	No	1
<i>hint usage</i>					
Emotion Labeling	Achievement Emotions & Metacognition	Students are asked to generally evaluate their mood using a multiple choice question: “How are you feeling right now? happy; frustrated; relieved; still confused”	No	Yes	1
Confidence Judgements	Metacognition	Students are asked to evaluate their confidence in solving problems upon seeing a problem, but before they are able to submit a response. This occurs three times during the learning activity on three separate problems	No	Yes	2

when the interventions were administered during the activity, and the number of problem sets in which the experiment was conducted. In the *embracing mistakes* treatment conditions, students received either *image* or *video* messages encouraging them to adopt a growth mindset by reappraising their mistakes as part of the learning process. The *inspirational quotes* intervention encouraged students to adopt joyful or hopeful emotions as they progressed through the activity by providing motivational messages and positive quotes from celebrities (Albert Einstein, Michael Phelps, and Nicki Minaj). The *social comparison* included two treatment conditions which presented students with normative information about the number of problems that students completed before mastery (*performance*) or the percentage of students who used hints during the activity (*hint usage*). The *emotion labeling* intervention engaged students in a metacognitive reflection on their emotions after they completed the first two problems during the activity. Similarly, the *confidence assessment* intervention encouraged students toward metacognitive reflection, but instead had students reflect on their confidence in solving prior to completing each problem in the activity. Notably, this final intervention was conducted in an experiment that did not include a post-test.

2.4.2 Data

The data included in our analyses were collected during seven school years in the United States (October 2015 through September 2022). During this time, the assignments were made available to teachers in middle school who used ASSISTments as an instructional tool and assigned these activities to their students as part of their lesson plans. During the experiment, 20,472 students worked on the experimental problem sets. Of the total sample, 3,722 students participated in multiple experiments. As students were randomized prior to participation in each experiment, this overlap does not bias the estimates, so all students were included in the analyses. These students participated in a total of 25,220 experimental mastery-based learning activities. The samples for each experiment are shown in Table 2.2.

Table 2.2: Experimental Sample Sizes

	Treatment		Control	
	<i>n</i>	%	<i>n</i>	%
Embracing Mistakes	2649	66.16%	1355	33.84%
<i>Image</i>	1341	33.49%		
<i>Video</i>	1308	32.67%		
Inspirational Quotes	2935	44.54%	3655	55.46%
Social Comparison	544	67.58%	261	32.42%
<i>Performance</i>	268	33.29%		
<i>Hint Usage</i>	276	34.29%		
Emotion Labeling	3887	54.02%	3309	45.98%
Confidence Judgement	3306	49.90%	3301	50.10%

2.4.3 Analytic Approach

To understand the various ways in which each experiment could impact student learning and learning-related behaviors, we utilized five different outcome variables, which align with our research questions. Each variable provides a different perspective on student learning and the student learning process.

Effect sizes were estimated for each outcome using a series of regression models. We estimated a model for each experiment. Each model compared the treatment condition to the specific control condition asso-

ciated with that experiment. For experiments with multiple conditions, one model was run for the overall treatment effect and then two subsequent models were estimated for each individual treatment effect. To account for potential inflation of Type I errors due to multiple comparisons, we adjusted the p-values of the effect sizes for family-wise error across the outcomes using the Bonferroni-Holm procedure within each experiment [1].

Equation 8.1 is the basis used to estimate the treatment effects for each experiment on each outcome. $Treatment_i$ is a binary indicator for whether the student received the treatment. β_1 is the effect of the treatment on the outcome. Specifics of the models vary by each outcome and the details of each outcome and model are described below.

$$outcome_i = \beta_0 + \beta_1 Treatment_i + \epsilon_i \quad (2.1)$$

2.4.3.1 RQ1 – Impact on Response Time

Response time is an important indicator of learning mathematics as pausing before submitting responses is associated with higher learning outcomes [143, 66]. In theory, pausing allows students to participate in metacognitive reflection, which can help them consider what strategies to apply to the problem [195]. In ASSISTments, response time is the time between entering the problem and the first action, which includes submitting an answer or requesting a hint.

To evaluate the impact of each experimental condition on response time, we estimated linear regressions for each treatment condition. Response time was averaged across all the problems the student completed within each mastery-based learning activity. As response time does not have a normal distribution (skew = 65.46, kurtosis = 6058.54), we used the log response time in the model.

2.4.3.2 RQ2 – Impact on Hint Usage

ASSISTments allows students to request hints that are developed specifically for each problem in the mastery-based learning activity. The problems in the master learning activity have up to six hints, each providing incrementally more information about how to solve the problem. If students select the final hint, they receive an entire worked example of the problem and the answer. Access to hints increases the probability that students will get the next problem correct [288, 273].

To evaluate the impact of each experimental condition on students' hint usage, we estimated logistic regressions for each treatment condition. We regressed the treatment indicators on whether the student re-

requested and received at least one hint during the mastery-based learning activity. Accessing hints was treated as a binary indicator because of the non-normal distribution of the hints usage (skew = 5.05, kurtosis = 45.12) and only 32.48% of the students across all experiments utilized hints during the activities.

2.4.3.3 RQ3 – Impact on Knowledge Component Mastery

For each mastery-based learning activity, students needed to reach a threshold of problems correct in a row to advance to the post-test. Thresholds ranged from 3 to 5 problems based on the experiment but were constant across conditions within each experiment. Completing the appropriate number of problems correct in a row indicates that students have mastered the knowledge component for the activity [179]. To assess the effects of each treatment condition on the likelihood that students will master the knowledge component of the activity, we employed logistic regressions.

2.4.3.4 RQ4 – Impact on Learning Performance

We used two outcomes to evaluate the treatment effects on learning performance: efficiency in mastery and post-test performance. Only students who mastered the knowledge component during the activity had efficiency metrics and were able to participate in the post-test. Students who did not master the knowledge component are considered part of the attrition group. The models for RQ3 serve as tests of attrition balance across conditions. Therefore, if the models from RQ3 show significant differences in the likelihood of mastery between treatment groups, the estimates of the RQ4 models predicting efficiency in mastery and post-test performance may be biased. Even if there is no evidence of attrition imbalance, the estimated effect sizes on efficiency in mastery and post-test performance are still limited to students who mastered the materials.

First, we evaluated how efficient students were in mastering activities' knowledge components. We used the number of problems each student took to master the knowledge component to create an efficiency variable. To ease interpretation, we standardized the number of problems to mastery and multiplied the variable by -1. Therefore, a positive value on the efficiency measure indicates fewer problems to mastery than a negative value, signaling greater efficiency. To evaluate the impact of each experimental condition on students' efficiency, we regressed the treatment indicators on the number of problems students completed to reach mastery.

Second, all but one of the experiments (*confidence judgments*) included post-test problems. The post-tests were brief, including only two or three items. The brevity of the post-test was due to implementation constraints of ecologically valid large-scale experiments, in which long post-tests are not feasible and would

likely result in greater attrition. Notably, this is not an ideal evaluation of individual students' learning, but provides an indication of the interventions' impact on learning across the entire population, in terms of differences in probability of answering a post-test problem correctly. The post-test problems were designed to be more difficult than mastery learning activities, but to utilize the same skills mastered within the learning activity. The post-test questions required students to transfer the skill learned during the assignment to complex problems often involving multi-step word problems. Each item is intended to assess how much learning occurred in the master-based learning activity.

To evaluate the impact of the treatments on the post-test, we estimated logistic regressions. Because aggregating a limited number of items does not produce a normal distribution, we regressed the treatment indicator on whether the student got each post-test problem correct. Equation 8.2 delineates this model. To account for variances in student ability, we included a random intercept for each student μ_i . To account for variations in post-test problem difficulty, we included a random intercept for each post-test problem μ_j . The post-test treatment effect is γ_{10} , which is the average difference in the log-odds of students in the treatment group answering a post-test problem correctly relative to students in the control group.

$$\text{logit}(\text{Student } i \text{ Gets Post-Test Problem } j \text{ Correct}) = \gamma_{00} + \gamma_{10}\text{intervention}_i + \mu_i + \mu_j \quad (2.2)$$

2.5 Results

Results from all of the models are displayed in Figure 2.1, which shows the effect size of each experiment on every outcome, including, confidence intervals, and statistical significance for each experiment on every outcome. The statistical significance indicated in the figure is based on the p-value corrected for family-wise error within each experiment. Notably, few of the interventions impacted students' learning or learning-related behaviors as measured by the studies' outcomes. The results for each experiment's impact are detailed in the sections below.

2.5.1 Embracing Mistakes Intervention

The embracing mistakes intervention did not significantly impact any of the outcomes. This finding was true regardless of whether the methods of message delivery (image or video) were evaluated individually or as one treatment. Overall, we found no evidence that messaging encouraging students to embrace their mistakes impacts their learning or learning-related behaviors.

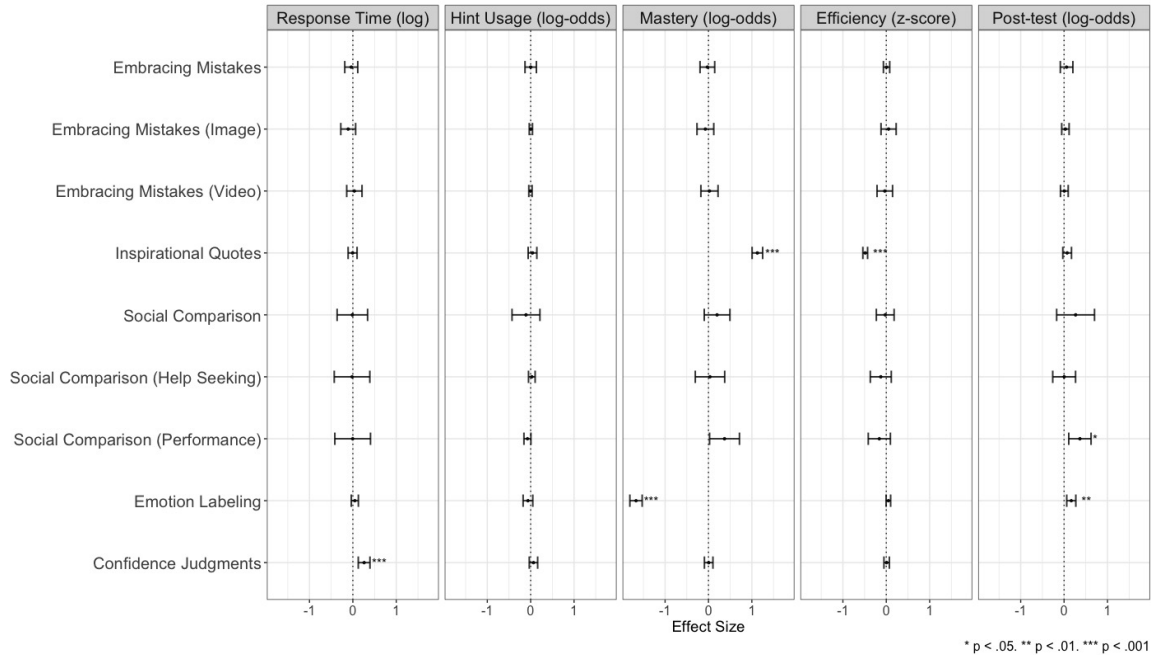


Figure 2.1: This plot provides the effect size, standard deviation, and statistical significance for each of the experimental conditions on each outcome. Statistical significance is based on p-values adjusted using the Bonferroni-Holm procedure within each experiment.

2.5.2 Inspirational Quotes Intervention

Although the inspirational quotes intervention had no statistically significant effect on response time, hint usage, or post-test performance, there were interesting patterns in the treatment’s effects on mastery rates and efficiency. Students who were exposed to inspirational quotes were more likely to master the activity’s knowledge component ($\beta_1 = 1.13$, $SE = 0.06$, $p < 0.001$). This effect size is substantial; students in the treatment condition had a .85 probability of mastering the knowledge component compared with .65 for the control group.

Of the students who did master the knowledge component, students in the treatment group were significantly less efficient than those in the control group ($\beta_1 = -0.49$, $SE = 0.03$, $p = 0.001$). Notably, the decrease in efficacy in the treatment group may be related to an increase in the number of students mastering. If the treatment inspired students who would otherwise give up to persist and complete more problems, ultimately mastering the knowledge component, this would drive up the average number of problems completed for the treatment group, thus explaining the decrease in efficiency.

Overall the treatment did not have a significant effect on student learning as measured by the post-test. However, attrition imbalance may have biased this result. Since only the students who mastered the activities’

knowledge component received the post-test, and the treatment encouraged students who would not have mastered to do so, this result may be biased towards the control group.

2.5.3 Social Comparison Intervention

Overall, there were no significant differences in learning or learning-related behaviors between the students who received the social comparison treatments and those in the control group. However, when the social comparison treatments were compared to the control individually, an interesting pattern emerged; the *performance* messaging seemed to have some impact on performance, whereas the hint usage messaging had no detectable impact on *hint usage*.

Students who received *performance* messaging outperformed the control on the post-test ($\beta_1 = 0.36$, SE = 0.13, $p = 0.022$). Notably, the *performance* messaging focused on efficiency by presenting the number of problems the average student completed before reaching mastery did not significantly affect students' efficiency. Yet, this condition had only a marginally non-significant effect on mastery; the effect size was moderate ($\beta_1 = 0.37$, SE = 0.17, $p = 0.141$), and this effect was non-significant only after the family-wise error correction and the majority of the confidence interval lies in the positive direction. This result suggests that fine-tuning the messaging may produce a better effect on mastery in future interventions.

The *hint usage* condition had a non-significant effect on learning or learning-related behaviors. Interestingly, this condition provided students with a message that the majority of students utilized hints, yet it had a non-significant effect on ($\beta_1 = 0.027$, SE = 0.04, $p = 0.999$).

2.5.4 Emotion Labeling Intervention

Emotion labeling had statistically significant effects on mastery, and the post-test but not on response time, hint, usage and efficiency. The effect on mastery was negative and substantial ($\beta_1 = -1.67$, SE = 0.07, $p < 0.001$). The probability of mastery for the treatment condition was .68, whereas the probability for the control was .92.

The effect on the post-test was significant and positive ($\beta_1 = 0.17$, SE = 0.05, $p = 0.006$). However, the effect may be caused by the attrition imbalance as fewer students in the treatment group mastered the knowledge component and had the opportunity to take the post-test. Therefore the causal effect of the treatment on the post-test is questionable. The pool of students who mastered the knowledge component may have been generally higher performing in the treatment than the control, thus biasing the results.

2.5.5 Confidence Judgments Intervention

The confidence judgments intervention did not significantly affect hint usage, mastery, or efficiency. The impact on response time was ($\beta_1 = 0.34$, $SE = 0.07$, $p < 0.001$). This suggests that the treatment condition may have slowed students' response time by encouraging them to think more about the problems. Yet, this increase in response time did not produce changes in other learning-related behaviors as measured by hint usage, mastery, or efficiency. The experiment did not include a post-test.

2.6 Discussion

The non-cognitive interventions analyzed in this paper produced mixed results regarding their impacts on learning and learning-related behaviors. Notably, none of the interventions produced consistently positive significant results across all the outcomes. The *social comparison* intervention, which provided normative messaging about student performance, was the only intervention to produce a positive result on the post-test while also passing the threshold of attrition imbalance by not having differences in mastery rates. Emotion labeling had negative impacts on both mastery rates and efficiency. Although students in this condition outperformed the control on the post-test, the reduced mastery rates in the treatment potentially biased this result, thus confounding the causal inference. Furthermore, *inspirational quotes* increased students' likelihood of mastery but failed to produce a positive effect on learning as measured by the post-test. Nevertheless, this may still be a favorable finding as students in treatment who would stop before mastering had they been in control may still have performed as well as control students. In sum, these experiments illustrated the difficulty of constructing non-cognitive interventions that positively impact students' behavior and performance on learning activities.

This difficulty in constructing impactful non-cognitive interventions may be due to the nature of non-cognitive factors. For example, for students to adopt a growth mindset while participating in a learning activity, they may need more extensive changes to their learning culture that transcend a small message at the beginning of an activity. One large-scale online mindset intervention found that the intervention's effects on math performance were dependent, in part, on whether the student's peer groups were supportive of a growth mindset [392]. Hence, a student who entered the *embracing mistakes* experiment may have already adopted a mindset based on their social learning environment and prior experiences that influences their perceptions of mistakes more than their exposure to the experimental messaging. Therefore, to impact student learning, non-cognitive intervention may require more robust changes to students learning environments than small messages and activities embedded in a CALP. Future work should examine whether CALPs can broadly

influence learning environments to affect non-cognitive factors and, subsequently, learning behaviors and outcomes.

One surprising finding was the nonsignificant effects of the *confidence judgments* intervention. This intervention was intended to inspire a metacognitive reflection on the alignment between students' abilities and problem difficulty. In theory, this should help students assess their knowledge state, which should ease the learning process [231]. In practice, this reflection did not affect the likelihood that students would master the knowledge component or improve their efficiency in learning. Notably, the other metacognitive intervention evaluated in this study, *emotion labeling*, had a negative impact on mastery and efficiency. These interventions may have been demotivating for students who either had low confidence in their ability through the activity or started with high confidence but had difficulty mastering the concept. There is evidence that metacognitive reflection could lead students to overconfidence [172]. The effects may differ based on students' levels of confidence or emotional state. Future analyses should focus on understanding the behaviors of these students based on their responses to the confidence judgment and emotion identification questions. Furthermore, qualitative data, such as information produced by 'think-alouds' and interviews, may help understand why emotion labeling can produce adverse effects.

Although our study provides interesting causal data on the impact of non-cognitive interventions in CALPs, there are some notable limitations. First, the lack of a multi-item post-test reduces the overall accuracy of assessing student learning and should be interpreted cautiously. It is infeasible to provide long multi-item post-test assessments in large-scale ecologically valid experiments, especially to estimate the effects of minor interventions embedded in short assignments. Future work should address this difficulty by developing valid and feasibly implemented measures in these contexts. Second, the implicit attrition caused by the master-learning activity potentially biases the results. Since only the students who master the activity get to take the post-test, our inferences about the impact of the intervention on learning as measured by the post-test is limited to a subset of the student population. Furthermore, when the intervention impacts the likelihood that students master the activity, this biases any estimates of effects on learning. Future work should consider using quasi-experimental methods to account for these imbalances or providing post-tests to all students regardless of mastery, perhaps at a set point in the activity as opposed to after mastery. Finally, our ability to understand the mechanisms of these interventions or make inferences about who may benefit from these interventions is limited by our lack of knowledge about the students who use many CALPs. More demographic information should be collected to understand the potential differential impact across populations.

2.7 Conclusion

Our work highlights the difficulty of developing non-cognitive interventions which positively impact students' academic behaviors and outcomes. Our findings confirm Yeager and Walton's assertion that non-cognitive interventions are not "magical" solutions but tools that only work in specific contexts with specific implementations [393]. Learning experience designers should note that applying non-cognitive theories while developing theoretically helpful features for students may result in adverse effects in practice. For example, we found that the emotional labeling intervention negatively affected students' likelihood of mastering knowledge components; hence, embedding meta-cognitive tasks into CALPS learning activities should be done with caution.

Our findings suggest that minor modifications in activity designs may be insufficient to change students' motivation and mindset in ways that impact behavior and learning. Designers may have to take global approaches to implement non-cognitive theories when building CALPs by considering how the entire program can be infused with motivational, mindset-oriented, and metacognitive content. Furthermore, simply changing the CALPs environment may not produce desired effects if the students' broader learning environments are not influenced (as found in [392]).

Finally, as we found both positive and negative results for these non-cognitive interventions, designers should always test the impact as they develop new features or more global meta-cognitive changes to programs. This process will ensure alignment between theory and practice, resulting in positive outcomes for learners.

MULTIPLE CHOICE VS. FILL-IN PROBLEMS: THE TRADE-OFF BETWEEN SCALABILITY AND LEARNING

Learning experience designers consistently balance the trade-off between open and close-ended activities. The growth and scalability of Computer Aided Learning Platforms (CALPs) have only magnified the importance of these design trade-offs. CALPs at scale often utilize close-ended activities (*i.e.* Multiple-Choice Questions [MCQs]) due to feasibility constraints of open-ended activities. MCQs unlike open-ended activities allow for easier development of immediate feedback. Our current study examines the effectiveness of Fill-In problems as an alternative to MCQs for middle school mathematics. We report on our experiment conducted from 2017 to 2022, encapsulating a total of 6,768 students from middle schools across the US. We observe that, on average, Fill-In problems lead to better post-test performance than MCQs; albeit deeper explorations indicate differences between the two design paradigms to be more nuanced. We find evidence that students with higher math knowledge benefit more from Fill-In problems than those with lower math knowledge.

Proper citation for this chapter is as follows:

Gurung, A. (In preparation). Multiple Choice vs. Fill-In Problems: The Trade-off Between Scalability and Learning.

3.1 Introduction

The rapid growth in technology and the ability to produce educational material accessible by enumerable learners has led to the development and adoption of Computer Aided Learning Platforms (CALPs) across educational sectors. With access to the internet, learners across the world use CALPs in the form of Learning Management Systems (LMSs), Massive Open Online Courses (MOOCs), and standalone online learning platforms. The past two decades have seen a drastic rise in the implementation and utilization of online educational materials [173, 136]. The promise of CALPs to transform education by scaling educational content by creating nearly ubiquitous access to responsive systems providing personalized instructions addressing individual learners' needs [43, 69] was met with great excitement. Yet today, many consider the promise of CALPs unfilled, citing lack of equity in usage and lower than expected impact [301, 389]. In order for

educational technologies to be impactful, they must be effective. So, designers of the learning and assessment activities, *i.e.*, Learning Experience (LX) designers, must ensure that their learning and assessment activities are optimized for scalability and effectiveness.

One of the fundamental challenges LX designers struggle with is balancing assignments' assessment and instructional value. Generally, LX designers have two options for problem types: open-end and closed-end. Designers can leverage the various affordances of CALPs to facilitate on-demand access to help and provide formative assessment; the same affordance is often harder to implement when designing open-ended activities. Open-ended activities, *e.g.*, traditional long/short answer questions (ORPs), and Fill-In problems require the learners to demonstrate their knowledge of the concepts in writing. In contrast, close-ended activities, *e.g.*, multiple-choice, ordering, and checking all that apply problems, provide options as a limited set of possible answer(s).

Open-ended problems often present feasibility constraints due to the requisite implementation resource they require, such as graders, physical space, materials, and technical and nontechnical support. In comparison, close-ended activities are easier to implement at scale due to their limited possibilities for answers, which makes them more suitable for automated grading, and feedback generation. As such, CALPs have widely adopted using close-ended problems, especially Multiple Choice Questions (MCQs). MCQs are used as auto-gradable activities in LMSs, *e.g.*, Canvas [174], GradeScope [339], Moodle [104], Blackboard Learn [12] and Schoology [285], MOOCs [158, 177, 197, 321]. Many other CALPs, *e.g.*, Daylite [38], Cognitive tutor [307], ASSISTments [155], and PeerWise [92], use MCQs to go beyond grading by providing automated feedback [140] and scaffolding to the address learner needs.

Although MCQs and other close-ended questions may be easily implemented at scale, LX designers and instructors also need to weigh the learning and assessment value of MCQs over ORPs. Both open and close-ended activities are designed to test learners' critical thinking and recall skills. However, some domain experts have reported perceiving open-ended activities to be more accurate because close-ended activities are susceptible to shallow learning and formulation of answers through recognition and synthesis [371]. A well-designed close-ended activity can be equally effective at facilitating critical thinking [372]. For example, a well-designed MCQ option addressing common misconceptions can be more effective at facilitating learning than an ORP – it can be challenging for LX designers and instructors to come up with such options every time.

Prior research has found mixed evidence on the effectiveness of close-ended activities over open-ended activities as a learning feature. When comparing MCQs to traditional open-ended problems, some reported on

the open-ended being more beneficial [3, 219, 75], others have reported on the benefits of MCQs [128, 350], while others have reported no difference between the two [214, 213, 345, 371]. While we do not dispute the advantage of using MCQs over ORPs at scale, we posit that such advantages are not as noticeable when considering Fill-In problems instead of ORPs. Fill-In problems by design, like MCQs, have limited answers and can be automated.

MCQs, in general, may have obvious scalability advantages to open-response questions; some LX systems use Fill-In problems, a type of open-ended problem, as an alternative to MCQs [155]. Fill-In problems with a limited set of possible answers can provide similar automation advantages to MCQs in on-demand help and formative assessment. Especially in mathematics, where the number of possible correct answers is limited and can be easily predicted. Fill-In problems can be implemented with similar affordances as MCQs while avoiding the risk of recognition and synthesis attributed to MCQs.

In this paper, we compare the effectiveness of MCQs against Fill-In problems by running an *en-vivo* study within a CALP where students worked on two separate assignments in mathematics. Both assignments were designed using Common Core State Standard [251]. Both assignments had MCQs and a Fill-In condition, and every student working on the activity was randomly assigned to one of the conditions followed by a post-test. This study was deployed in Fall 2017 and is still running to date. Our analysis spans five school years and four summers in the United States. The primary contribution of this paper is to explore the difference in learning value between MCQs and Fill-In problems.

Prior works by Wang *et al.* [371] showed empirical evidence demonstrating MCQs and ORPs to be similar at assessing students. Through this work, we aim to complement Wang *et al.* [371] by exploring the learning outcomes of assignments by shifting the focus from assessment to learning. We posit that differences in learning outcomes between MCQs and Fill-In problems exist. As such, this paper has three goals: (a) to explore the differences in learning outcomes between MCQs and Fill-In problems, (b) to examine the effect of the atypical learning contexts, pandemic, and summer sessions, on the learning outcomes, and (c) to investigate the personalization effects in learning outcomes between MCQs and Fill-In problems. The following paragraphs briefly describe our approach to pursuing the three goals of this paper.

First, we analyze the effectiveness by comparing MCQs with Fill-In problems for students working on mastery-based assignments. We utilize an experimental design to test the effect of problem type on a post-test administered upon mastery. Analyzing differences in post-test performance between students in MCQs and Fill-In conditions allows us to estimate the relative learning value of each problem type. We find that, on average, students in the Fill-In problem are more likely to answer the post-test correctly than students in the

MCQ condition. We also find evidence of heterogeneous effects across problem sets.

Second, we examined the effectiveness of the MCQs and Fill-In problems when the learning context changed. We examined two periods in which students were likely to experience education outside of traditional in-person classrooms: during the Coronavirus pandemic (COVID-19) and during typical summer months in the United States (summer break). We observed a significant average decline in student performance for the post-tests during the COVID-19 Period; however, students in the Fill-In problem still performed better on the post-test. We did not detect any significant differences in the student performance on the post-test during the summer break.

Third, we expanded upon our findings and examined differences in the impact of the Fill-In problems on the post-test performance of the students based on students' math knowledge prior to the experiment. We found that the Fill-In design was more effective at helping students with higher prior performance, with no difference in student performance between the two problem types for students with average prior performance. However, we find evidence that students with the lowest prior performance benefit more from MCQs, which suggests that the effect of problem types depends upon the learners' knowledge level. LX designers and instructors may need to personalize their assignments by accounting for prior math knowledge to be more effective at impacting the learning outcome of their students.

3.2 Prior Works

Over the years, various prior research has explored the efficacy of utilizing MCQs over ORPs and found mixed results. Researchers have explored the feasibility of the two design approaches and have shown ORPs to be more costly towards instructor resources than MCQs [32]. It is critical to take the situational context in which MCQs and ORPs are implemented, as each has unique advantages and disadvantages. For example, Polat [284] cites studies [191, 300] that detail the benefit of MCQs when mass-testing students, *i.e.* SAT, TOEFL, as they are easier to grade. Research conducted in STEM courses has shown that students perform better on MCQs than ORPs [128, 350]. Though MCQs are prominently used by CALPs, a major drawback of MCQs is the fact that there is an element of synthesis, guessing, and recognition of distractors¹ [284, 150, 59] which can be misconstrued as knowledge or lead to shallow learning. As such, MCQs have the possibility of decreasing the reliability and validity of the tests [84]. Ruit & Carr [317] detail that distractors can jog learner memories on problems they are trying to solve, thus decreasing the reliability of the MCQs for measuring student knowledge. Additionally, MCQs cannot measure creative thinking and idea generation [52]; skills

¹Distractors, sometimes referred to as "Lures" in other academic domains, are incorrect answers to a multiple-choice question that distract students from the correct answer by providing erroneous information.

directly addressed by ORPs. ORPs can exhibit a student’s high-level thinking and reasoning on a given problem while eliminating the element of guessing that is typically seen in MCQs [284].

In a recent publication at CHI, Wang *et al.* [371] focused their attention on the assessment value of MCQs vs. ORPs while teaching Human-Computer Interaction (HCI) and found there to be no difference between the “assessment value” of MCQs and ORPs despite various domain experts intuiting ORPs to have more value. The domain experts, HCI instructors, attributed ORPs to improving idea generation, information recall, and critical thinking skills while expressing concerns about the risk of recognition and synthesis in MCQs. Wang *et al.* present empirical evidence demonstrating domain experts’ blind spots in determining the problem difficulty of MCQs compared to ORPs – insight that can be valuable to assessment designers. While we do not dispute the findings of Wang *et al.* [371], we are concerned that LX designers and instructors can misinterpret their work to justify the usage of MCQ in assignments that focus on facilitating learning. We would argue that the primary benefit of assignments is their “learning value” and the “assessment value” only plays a significant role in a few scenarios where instructors wish to assess or test the learner’s knowledge categorically.

3.2.1 *Mastery-Based Learning Activities*

Over the years, pedagogical research has seen a growing trend toward adapting mastery-based learning. Rather than assuming learning upon completing a certain number of hours on the material, mastery-based learning [344] requires learners to demonstrate knowledge and skill in the concepts before progressing to the next topic. Mastery-based learning approaches have shown to reduce variance in student aptitude [11, 196], increase long-term retention of knowledge [196], change student attitude towards content [11, 196], and increase self-belief [11, 145]. It is important to note that mastery-based learning is not without risk. If left unmoderated, mastery-based assignments can lead to overworking students, causing frustration and adversely affecting students’ perception of the material and their abilities.

One of the primary features of mastery-based learning is to provide students with the ability to practice the skills that allow the teachers to assess their students’ abilities while facilitating learning opportunities. CALPs, by design, have an advantage when implementing mastery-based assignments, as automation is a fundamental property of computers. Various CALPs have explored the implementation of mastery-based assignments using various approaches. While some platforms, such as Khan Academy [255, 194], and ASSISTments [155], have explored using an arbitrary threshold of N-Consecutive Correct Responses (N-CCR), others have relied on more precise measures of mastery using Knowledge Tracing (KT) models [79]. KT

models predict student performance in future problems by leveraging their past performance on similar or related skills. Both N-CCR and KT approaches have their merits and flaws; N-CCR is more explainable and controllable by teachers, whereas KT models are harder to understand for the teachers but more accurate at estimating mastery. While a simple heuristic of N-CCR could be considered rather simplistic, Kelly *et al.* [180] found an N-CCR design, with $N = 3$, to compare mastery learning between KT models and the N-CCR approach. Prihar *et al.* [292] have reported on experiments extending the N-CCR experiments exploring the benefits of using $N = 2, 4$, and 5 as a threshold and found $N = 3$ to be an optimal threshold in the domain of learning mathematics. While a simple N-CCR design is easy to implement and interpret, some have expressed concerns about the risks of inequitable outcomes due to the N-CCR design's assumption about student learning [180, 170] that may not hold in practice across all contexts [166, 102, 101].

3.3 Study Design

We designed the experiment within a CALP system [155] to evaluate the impact of problem types; MCQ and Fill-In problems. To conduct this experiment, we created two problem sets teaching two different mathematical concepts - Greatest Common Factor (GCF) and Evaluating Expressions (EE). Both the problem sets were designed using the Common Core State Standards [251] in which the GCF problem set used the grade six curricula, whereas the EE problem set used the grade seven curricula.

As seen in Figure 3.1, each problem set had a mastery learning component and a post-test. In the mastery learning component, students were randomized across two conditions; MCQs vs. Fill-In problems. Both conditions require the students to answer three consecutive problems correctly to demonstrate mastery over the material. If a student submits an incorrect response on their first attempt, the consecutive correctness counter gets reset to 0. There is a daily limit of 10 problems per condition unless the student correctly answers the 9th or 10th problem; in such cases, the student can attempt up to 11 or 12 problems to demonstrate mastery. If the students cannot demonstrate mastery within the first 10 problems, they must wait until the next day to work on the problem set. While working through the problems in the assignment, students can ask for help in the form of hints where the bottom-out hint has the answer to the problem. Students only get full credit if they answer the problem correctly on their first attempt without asking for any help.

Upon acquiring mastery, students answered two post-test problems. These problems are more complex Fill-In problems than those in the mastery learning component on the same topic as the experiment. Examples of the problems in the experiment (MCQs vs. Fill-In problems) and the post-test are provided in Figure 3.2. These examples display the relative complexity of the post-test compared to the problems in the mastery

learning component. The post-test problems were intentionally designed to be more complex than the problems in the mastery component such that we can evaluate the transfer of knowledge from the mastery learning component to more complex situations, like complex word problems. We chose to utilize Fill-In questions because the requirement that the student produces answers independently allows us to measure knowledge transfer more effectively. Thus, this post-test serves as an estimate of the learning benefit from the mastery learning component.

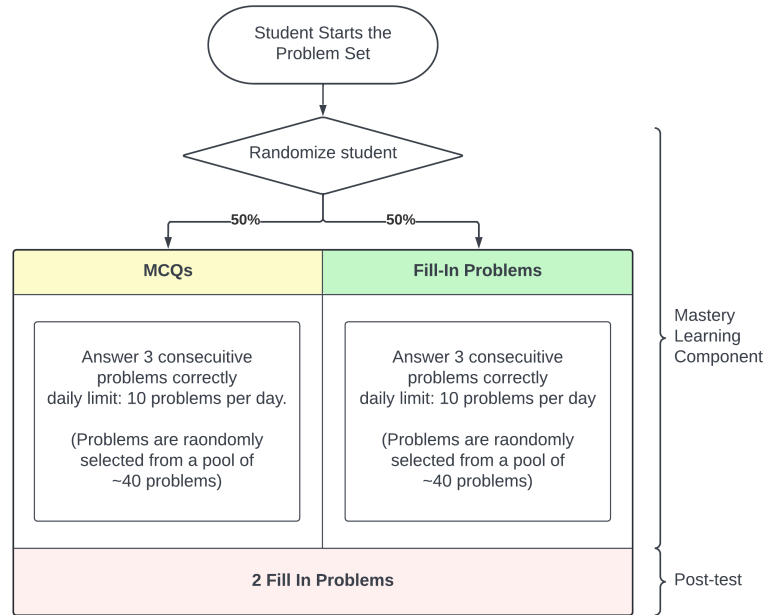


Figure 3.1: Breakdown of the experimental design. It has two conditions where students are assigned a mastery based assignment with MCQs or Fill-In problems. Upon mastering the content students are asked to answer two Fill-In problems that with higher difficulty than the problems in the experiment.

3.3.1 Description of Dataset

The experiment data was collected across five school years in the United States (2017-18, 2018-19, 2019-20, 2020-21, 2021-22). During this time, the assignments were made available to teachers in middle school who use the ASSISTments platform as an instructional tool and assigned these problem sets to their students as part of their lesson plans. During our study, 192 teachers assigned 430 problem sets in 383 classes. The assignment-to-class relationship is not one-to-one because some teachers prefer to divide their classes into smaller groups and assign the same assignment to each group separately. During the experiment, 6,774 students started the experimental problem sets. A small number of students (20), worked on both problem sets. In such instances, we only include the student data from the first experiment they participated in and

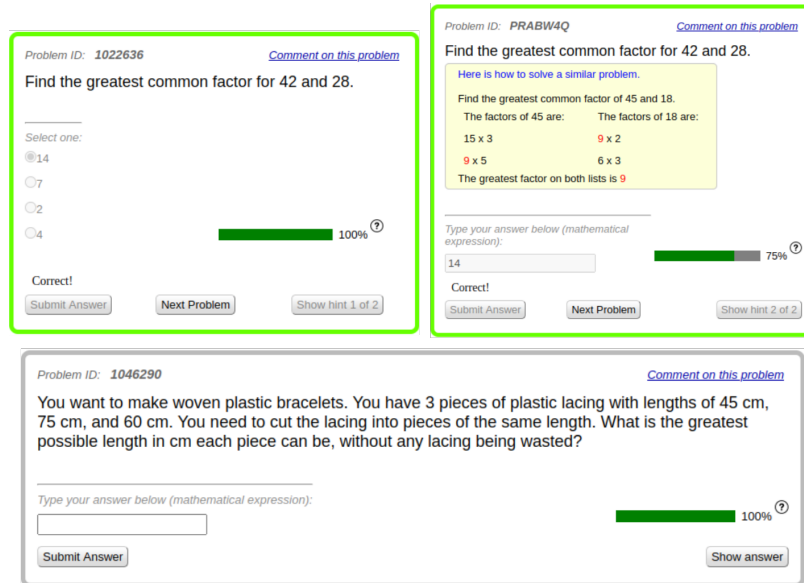


Figure 3.2: Example Problems from “Problem Set 1: Greatest Common Factor”

drop the other records.

The dataset includes information on the treatment condition (MCQ and Fill-In problems), the problem set (MCQs and EE) the students worked on, their correctness on all the problems during treatment, their correctness on the post-test problems, and when each post-test problem was completed. The dataset also contains information on students’ prior percent correct on problems in the ASSISTments platform prior to participating in the experiment, along with unique identifiers for teachers and classes level information.

3.3.2 Attrition

Table 3.1 displays the experiments’ attrition rates and balance test statistics. Overall the attrition rate was 27.07%. Attrition occurred at one of three levels. First, 18.02% of students did not exhibit mastery in the learning component and, as such, could not take the post-test. Second, 9.91% of students exhibited mastery but did not start the post-test. Third, some students completed only one of the post-test problems. These students were not excluded from the analysis and, therefore, are not included in the overall attrition.

To ensure that attrition does not bias our estimates of the treatment effect, we tested attrition balance across conditions in two ways. First, we took the raw difference between attrition for each condition and compared it against the U.S. Institute of Education Sciences (IES) [169] standards. We also conducted chi-squared homogeneity tests to evaluate whether these differences in attrition were statistically significant. We conducted these analyses at each level of attrition. Based on the results of both IES recommendation and

Table 3.1: Student Level Attrition

	<i>All</i>	<i>Fill-In</i>	<i>MCQ</i>	$ Difference $	<i>IES Threshold</i>	χ^2	<i>p</i>
Overall attrition	27.07%	27.02%	27.13%	0.11%	5.40%	1.24	0.266
Did not reach mastery	18.02%	18.93%	17.18%	1.75%	5.70%	3.37	0.066
Did not start post-test	9.05%	8.09%	9.95%	1.86%	6.00%	1.44	0.231
Did not complete post-test	8.45%	7.92%	8.93%	1.01%	6.30%	1.72	0.190

Table 3.2: Descriptive Statistics

	All		Fill-In		MCQ	
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>
Average problems to mastery	4.94	2.99	4.92	2.81	4.97	3.16
Average hints access	0.58	1.88	1.04	2.55	0.15	0.56
Average time to first response (sec)	61.32	457.61	74.61	417.29	48.76	492.42
Average score on mastery component	86.21	16.98	88.80	13.81	83.77	19.15

chi-squared test, attrition was balanced across conditions at all levels.

3.3.3 Descriptive Statistics

Descriptive statistics on student behavior in the mastery learning component are reported in Table 3.2. Although there were no substantial differences in the average number of problems a student took to reach mastery across conditions, there are notable differences in the other behaviors. Students in the Fill-In condition accessed more hints, took more time before submitting their first response, and had higher scores on the problems in the mastery learning component.

3.4 Analysis 1: Impact of Fill-In Problems on Student Learning

In this section, we evaluate the effect of the Fill-In problems, relative to MCQs, in helping students learn the underlying concepts addressed by the problem sets. We hypothesize that Fill-In problem sets have a positive impact on learning and, therefore, students who work through Fill-In problems will perform better than students who work through MCQs on the post-test. Our hypothesis is based on the assumption that

students in the Fill-In condition are more likely to exert effort towards understanding the underlying concepts as they must independently provide the answer, *i.e.*, critical thinking and recall, without the ability to access possible answers. Alternatively, students who work through MCQs can be more susceptible to shallow learning through educated guesses and deduction of the correct answers, *i.e.*, recognition and synthesis, without completely grasping the mathematical concepts addressed by the content. Therefore, we expect students in the Fill-In condition to be more likely to answer the post-test problems correctly than those in the MCQ condition.

3.4.1 Methods

To evaluate whether students are more likely to learn while working on Fill-In problems than on MCQs, we estimated a series of mixed-effect logistic regressions using the `lme4` package in R [33]. For each of our models, we regress indicators on whether the students got each individual post-test problem correct on the first attempt using a binary indicator for the Fill-In problem set condition. We use this method because averaging the post-test problem together would not have created a continuous variable, as there were only two post-test problems for each problem set. Therefore, a linear regression likely has a poor model fit. By using a logistic regression model, we can treat each post-test problem individually – this allows us to include students even if they did not complete both post-test problems. The inclusion of these students will not bias our effect estimates as completion of the post-test was consistent across conditions (see Section 3.3.2).

For our first model, Model 1, we include random intercepts for post-test problems to account for differences in problem difficulty and students because students completed multiple post-test problems. We also include random intercepts for the students' classes because their classroom context could influence their learning behaviors, and students are often grouped within classes with students of similar ability.

For any given post-test question j completed by student i , the equation for the likelihood of correctness is (8.1) where γ_{00} is the fixed intercept, μ_{0i} is the random intercept for each student, μ_{0c} is the random intercept for each student's class, and μ_{0j} is the random intercepts for each post-test problem. Fill-In_i is a binary indicator for whether a student is in the Fill-In condition, and the coefficient for the effect of the Fill-In problem sets condition is γ_{10} .

$$\text{logit}(\text{Student } i \text{ Gets Post-Test Problem } j \text{ Correct}) = \gamma_{00} + \gamma_{10}\text{Fill-In}_i + \mu_{0i} + \mu_{0c} + \mu_{0j} \quad (3.1)$$

Model 1 allows us to determine whether working through Fill-In problems increases students' likelihood of learning the content based on their post-test performance. Since the experiment included two separate

problem sets, completed by different sets of students, we also ran two more models estimating the effect of Fill-In problem sets on learning when learning GCF (Model 2) and EE (Model 3) separately. Analyzing GCF and EE on their own allows us to determine whether the effect exists across different content areas.

3.4.2 Results

Model 1 shows a positive causal effect of Fill-In problems on student learning (Table 3.3). Overall, students who worked through Fill-In problem sets were significantly more likely submit correct responses on the post-test than those who worked through MCQs ($\gamma_{10} = 0.23$, $SE = 0.06$, $p < 0.001$). The likelihood of answering problems correctly on the post-test was true even after accounting for the variance explained by the individual student, their class, and the problem.

We also evaluate the variance in post-test performance associated with the students and their classes by examining the variances of μ_{0i} , τ_0 . Much of the variance in performance was associated with the student ($\tau_{0i} = .83$). Yet the students' classes still accounted for a substantial proportion of the variance ($\tau_{0c} = .60$), suggesting that the students learning environment is an important factor. We explore this finding further when we examine the potential influence of students learning environment in Section 3.5.

Although the overall effect of Fill-In problem sets on post-test performance was positive and significant, our subsequent models showed that the impact of Fill-In problems on learning may be more nuanced. Models 2 and 3 show the effect is only significant when the students were in the GCF problem set (Model 2, $\gamma_{10} = 0.26$, $SE = 0.07$, $p < 0.001$) and not significant when students were in the EE problem set (Model 3, $\gamma_{10} = 0.14$, $SE = -1.60$, $p = .$). It is also notable that the intercepts of the two problem sets differed as well, *i.e.*, the average likelihood of the students getting the post-test problem correct is higher for students in the GCF problem set (Model 2, $\gamma_{01} = -0.55$, $SE = 0.24$, $p = 0.024$) than for students in the EE problem set (Model 2, $\gamma_{01} = -1.60$, $SE = 0.14$, $p = 0.016$).

Due to the non-significant finding for the EE problem set, we ran a post-hoc model² to test whether the effects of Fill-In differed by condition. We did this by adding an interaction between a binary indicator for EE and Fill-In problem sets to Model 1. The estimate for this interaction was non-significant $\gamma_{03} = -0.14$, $SE = 0.12$, $p = 0.264$). While model 3 indicates that the effect of Fill-In on post-test performance is not significant on its when the EE problem set is isolated. However, our supplemental model shows that the effects of fill-in conditions between GCF and EE are not significantly different, *i.e.*, the 95% confidence intervals across problem sets overlap.

²The full model output is available in the supplemental materials.

Table 3.3: Analysis 1 Results

<i>Predictors</i>	Model 1 (GCF & EE)			Model 2 (GCF)			Model 3 (EE)		
	<i>Log-Odds</i>	<i>SE</i>	<i>p</i>	<i>Log-Odds</i>	<i>SE</i>	<i>p</i>	<i>Log-Odds</i>	<i>SE</i>	<i>p</i>
Intercept	-1.02	0.27	< 0.001	-0.55	0.24	0.024	-1.60	0.16	< 0.001
Fill-In	0.23	0.06	< 0.001	0.26	0.07	< 0.001	0.14	0.14	0.216
Random Effects									
σ^2	3.29			3.29			3.29		
τ_{00}	0.83 _i			0.73 _i			1.21 _c		
	0.60 _c			0.45 _c			0.99 _i		
	0.25 _j			0.11 _j			0.01 _j		
ICC	0.34			0.28			0.40		
N	4940 _i			3362 _i			1362 _i		
	363 _c			266 _c			60 _c		
	4 _j			2 _j			2 _j		
Observations	9657			6561			3096		

Taken as a whole, these findings show that there is a positive average effect of Fill-In problems on post-test performance in the GCF problem set. The models suggest that the effect was slightly smaller, but still positive, in the EE problem set, but the data are consistent with other possibilities, including similar effects across problem sets, or zero or negative effects in the EE problem set.

3.5 Analysis 2: Effect of Fill-In Problems During Atypical Learning Periods

Our analysis in Section 3.4.2 revealed a positive effect for Fill-In on learning, but also some differences in these effects based on the content of the problem sets. Now we seek to explore whether there are some other nuanced differences in effects based on the differences in students' learning environments. Differences in learning environments could impact how students interact with the programs. For example, a student completing problem sets in a CALP at home may be more likely to utilize support outside of the CALP, such as asking a parent for help or using an online search engine to find the correct answer for the mastery-based problem set and the post-test problems. Alternatively, if a student is working in a classroom, they may not have the same freedom to use these external supports while gaining access to teachers and tutors. The variation in learning environments could cause the effect of the Fill-In problem sets to vary as well.

In order to test whether the effect is consistent across learning environments, we tested whether the effect of Fill-In problem sets on learning differed across two time periods where students worked in an atypical learning environment. For the sake of simplicity, we define business as usual in-person learning during a regular academic year in the US as a typical learning environment. A typical academic year in the US is 9 months long and usually begins in mid to late August and ends in late May to early June. First, we examine problems completed during the height COVID-19 pandemic in the US - from March 2020 to June 2021 - in which students were likely to work remotely for at least some of their school time [120, 351]. Next, we examine problems completed during the typical US summer break period - June to August - in which students are also likely to be either at home or in summer school. Although we do not have information about each individual student's learning environment, this examination intends to provide insight into whether the effects vary on average as the student populations' learning environment shifted during these atypical learning time periods.

We suspect that students will be more likely to use external supports that would provide them with answers regardless of condition when in atypical learning environments. Therefore, we hypothesize that the effect of the Fill-In problem sets on learning will be diminished during both the height of the COVID-19 pandemic in the US and the typical US summer break months.

3.5.1 *Methods*

To evaluate our hypothesis, we built upon Model 1 (see 8.1), by adding interactions in Model 4 between the Fill-In condition variable and dummy variables, indicating whether the problem sets were completed during each atypical learning period (Pandemic and Summer Break). These interactions allow us to evaluate whether the effect of Fill-In problem sets on learning differed during each atypical learning period relative to the typical school year. We also allowed the effect of the Fill-In condition to vary for each class by adding in a random effect at the class level, γ_{11} ; allowing us to qualify how much the effect of the Fill-In condition differed by the class environment.

3.5.2 *Results*

The model estimates for Analysis 2 are in Table 3.4. The main effect of the Fill-Answer is significant and positive in both Model 4 ($\gamma_{10} = 0.24$, SE = 0.08, $p = 0.004$) – indicating that in typical learning periods, the Fill-In problem increases the likelihood of correct response on the post-test problems. Interaction coefficients are not significant for either the Pandemic period ($\gamma_{40} = 0.0003$, SE = 0.11, $p = 0.994$) or the Summer Break

Table 3.4: Analysis 2 Results

<i>Predictors</i>	Model 4		
	<i>Log-Odds</i>	<i>SE</i>	<i>p</i>
Intercept	-0.87	0.27	0.001
Fill-In	0.24	0.08	0.004
Pandemic	-0.27	0.12	0.027
Summer Break	0.38	0.27	0.164
Fill-In x Pandemic	0.003	0.11	0.994
Fill-In x Summer Break	-0.12	0.28	0.681
Random Effects			
σ^2	3.29		
τ_{00}	0.83 _c		
	0.58 _c		
	0.25 _j		
τ_{11}	0.0003 _c		
ICC	0.33		
N	4940 _i		
	383 _c		
	4 _j		
Observations	9657		

period ($\gamma_{50} = -0.12$, $SE = 0.28$, $p = 0.681$). Therefore, the effect of the Fill-In problem sets on learning persists regardless of the learning environment. The variance of γ_{11} (τ_{11}) was small and Model 5 (0.0003), showing that the effect of Fill-In condition did not vary greatly by class. Contrary to our expectation, the results from both models show no evidence that the students learning environment influences the effect of Fill-In problems on post-test performance.

3.6 Analysis 3: Personalization Effect of Fill-In Problems

Finally, we further evaluated the robustness and nuances of the Fill-In problem effect on student learning by examining whether the effect differed based on students' math ability. Students may benefit differently

from problem types based on the level of their math knowledge. Perhaps higher-knowledge students benefit from practicing critical thinking and recall while completing Fill-In problems without the options provided in MCQs. In contrast, lower-knowledge students might benefit from seeing possible answers that help them develop their intuition and understanding of the material. Alternatively, lower-knowledge students might be deceived by spurious MCQ options and, therefore, might benefit from having to work through those problems without the options provided in the MCQs. We hypothesize that students' prior knowledge will influence the effectiveness of the two treatment conditions, but we do not have any intuition as to the direction of this interaction as we view both scenarios as equally likely.

3.6.1 *Methods*

To determine whether the effect of Fill-In problems differs based on students' general knowledge of math concepts, we added an interaction between students' prior performance in the ASSISTments platform and the Fill-In Condition. For students who completed problems in the ASSISTments platform prior to working on the experiment, we have data on their prior performance. We use students' average score in these problems as an estimate of students' math ability. Prior performance was added as a standardized score to the model to improve interoperability. The standardization was calculated using grand mean centering (using the mean and standard deviation of the entire sample) instead of within their classes because we wanted to estimate the effect of Fill-In problems for the total population average as opposed to the within classes.

Of the original sample, 1,643 students had completed at least ten (10) problems in the CALPs before the experiment. We excluded students with fewer than ten prior problems completed in the ASSISTments platform we used in this study prior to the experiment as this amount of data would provide poor estimates of math ability. This exclusion was balanced according to the conditions: 19.96% students from the Fill-In condition and 19.72% students from the MCQ condition. Therefore, the exclusion does not bias our estimates of the Fill-In effect. Prior performance of this analytic sample ranged from 14.29% to 100% with a mean of 70.05% and a deviation of 15.02%.

3.6.2 *Results*

Table 3.5 displays the results for Model 5. The main effect (γ_{10}) is the effect of the Fill-In problem for the students who received the average score because scaled prior performance is centered at the mean. The effect of Fill-In on the likelihood of getting the post-test correct for students with average prior performance effect is non-significant ($\gamma_{10} = 0.13$, $SE = 0.09$, $p = 0.156$). The interaction between prior performance and

Table 3.5: Analysis 3 Results

<i>Predictors</i>	Model 5		
	<i>Log-Odds</i>	<i>SE</i>	<i>p</i>
Intercept	-3.45	0.24	0.001
Fill-In	0.13	0.09	0.156
Prior Performance (Z-score)	0.55	0.08	<0.001
Fill-In x Prior Performance (Z-score)	0.20	0.10	0.042
Random Effects			
σ^2	3.29		
τ_{00}	0.61 _i		
	0.77 _c		
	0.0 _j		
ICC	0.32		
N	1643 _i		
	100 _c		
	4 _j		
Observations	3253		

the Fill-In problem set is significant and positive ($\gamma_{30} = 0.20$, $SE = 0.10$, $p = 0.042$). Therefore, the effect of Fill In problems compared to MCQs appears to depend on the student’s prior math ability – especially for high-performing students, Fill-In problems led to better post-test performance, while for lower-performing students, the effect was smaller and possibly negative.

We visualize the interaction between Fill-In problems and prior performance in Figure 3.3 by plotting the predicted probability of a correct response on the post-test for each post-test attempt by Prior Performance for both Fill-In and MCQ conditions. The probabilities were predicted using Model 5. Notably, the visualization shows a negative effect of Fill-In problems for students with lower prior scores. To test whether this effect is significant, we ran a post-hoc ³ model based on Model 5 with prior performance low-end centered so that the main effect will be for the effect of Fill-Ins for students with the lowest prior performance scores. The main effect was not significantly significant ($\gamma_{10} = -0.61$, $SE = 0.38$, $p = 0.114$). In summary, we have strong

³The full model output is available in the supplemental materials.

evidence that the effect of Fill-In problems is greater for students with higher prior performance compared with students with lower prior performance. However, there is insufficient evidence that the effect of MCQs is negative for students with lower performance.

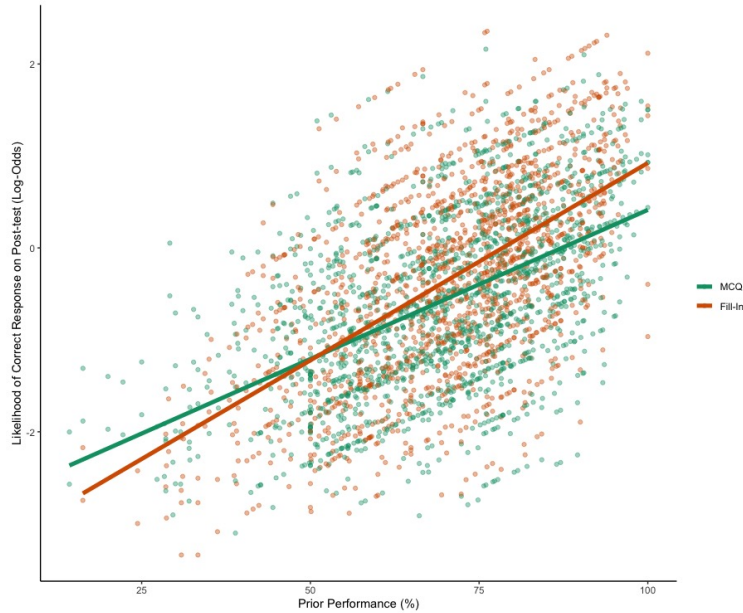


Figure 3.3: Interaction effect of problem type and prior performance on likelihood of post test correctness.

3.7 Discussion and Future Works

Our analysis found that, on average, problem sets with Fill-In problems lead to better learning outcomes than MCQs. Fill-In problems were still more beneficial than MCQs when we measured learning outcomes by segmenting the experiment into the pre-pandemic, pandemic, and summer sessions. The benefits of Fill-In problems over MCQs across various typical and atypical settings attests to the robustness of the effects of Fill-In problems. However, we observed that Fill-In problems' benefits have certain contextual constraints. The impact of Fill-In problems on learning was not significant with certain content and for certain students. Yet, notably we found no evidence that students who worked through MCQs ever significantly outperform those who worked through Fill-In problems. In sum, despite the nuance exposed by our analyses, Fill-In problems have an average positive effect on student learning compared with MCQs.

One interesting nuance was the significant effect of Fill-In problems on post-test performance for GCF and not EE. There are several contextual differences between the two problem sets, *e.g.*, mathematical concept, grade level, and problem complexity. The primary cause of the difference in the effectiveness of the two problem sets is not obvious as both were structurally the same in terms of on-demand help and formative

feedback. A potential aspect we look to explore in future works is the role of desirable difficulty [42] in the effectiveness of Fill-In problems. In Table 3.2, we reported the students' time on task while working on the problems in the mastery-based component. We observed that the students in the Fill-In problems invested more time formulating their answers than those in the MCQs because they perceived the problems to be more difficult. The difference in the time commitment could be a proxy for the students' perceived difficulty of the problems that consequently resulted in the difference in learning outcomes between conditions. A possible explanation for the underlying cause of the difference in learning outcome could be founded in the principles of learning theory exploring desirable difficulty [42].

Furthermore, our analysis in Section 3.6 exploring the personalization effects of Fill-In problems indicates that not all students benefit equally from Fill-In problems. We observed that students with higher prior performance benefited more than students with lower prior performance. Although this finding implies that students with lower prior performance might benefit more from MCQs than Fill-In problems; however, we cannot make more substantial claims due to the sparsity of students with low prior performance in the data. Despite this uncertainty, our analysis shows impact differentials for Fill-In based upon students' knowledge before beginning the activity. There are some plausible explanations for this phenomenon. High-knowledge students may have the ability to learn the concept addressed in the problem sets, but may need the challenge of having to produce the answers themselves without the MCQ options to truly benefit from the activity. Alternatively, lower-knowledge students may benefit from the options in the MCQs but are less likely to learn the concepts well enough to transfer their knowledge to problems where they must provide the answer independently. Regardless of the underlying mechanisms behind the penalization effect, the finding provides evidence that LX designers and instructors may have to consider adapting problem types to students' needs.

Building on our findings from this paper, we aim to further study the processes that cause Fill-In problems to impact student learning. Using multiple mediation analysis, we plan to evaluate potential paths that lead from problem type to mastery demonstration to discern and assess how problem types influence student behaviors which, ultimately, cause differences in learning outcomes.

There are a few key limitations of our work. First, we conducted experiments on two very specific content areas, where we found evidence that content may influence the effect of the problem type on learning. This research should be replicated using different content areas across different subjects to fully understand the heterogeneity of the impact the problem types can have on student learning. A further limitation of the current study is the lack of student demographic information. The ASSISTments platform we used in this study does not collect personally identifying information about the students, per the IRB Protocol; thus, we cannot make

any advances in understanding the more fine-grain differences within our sample.

Our work has an experimental design limitation as we only use Fill-In problems for our post-test. Concerns regarding this design limitation are valid; yet, we argue that knowledge, by nature, should be transferable upon mastery and, as such, would be independent of the instrument used during evaluation. While such assumptions regarding transferability can be problematic, the balanced post-test completion rates across conditions indicate that students from both conditions were comfortable with the design of the post-test. Although we feel that using the Fill-In problem is justifiable as Fill-In problems are an accurate measure of student ability. Further exploration using a combination of both MCQs and Fill-In problems would help establish the optimal approach in the design of assignments as the combination of both activities could enhance learning outcomes or, conversely, the switch between problems in the post-test could cause cognitive load leading to higher dropout rates. Similarly, additional work exploring the benefits and drawbacks of other types of close and open-ended activity design would be beneficial to understand their assessment and learning value.

3.8 Conclusion

At the onset of this research, we posited that there are differences in the learning outcomes when students engage with MCQs and Fill-In problems in CALPs. We observed that, on average, students had better learning outcomes when using mastery-based assignments with Fill-In problems over MCQs. We also demonstrated the robustness of our findings by evaluating them across various contextual scenarios, *i.e.*, pre-pandemic, pandemic, and summer sessions. We took a comprehensive approach and evaluated the personalization effects of the two methods, where we observed high-performing students benefiting more from Fill-In problems.

We consider our work to complement prior work by Wang *et al.* [371], which examined the assessment value of MCQs with traditional ORPs. Wang *et al.* [371] reported no difference between ORPs and MCQs in assessment value, whereas our work found a significant difference in the learning value of Fill-In problems over MCQs. We consider our research to be a significant contributor to the domain of learning experience design, as we provide evidence that problem types have an impact on learning outcomes. Our findings support the use of Fill-In problems for learning activities but also provide evidence that different students may require different types of activity design, to learn more effectively. We believe that LX designers and instructors will benefit from our findings when designing learning and assessment activities where they are continually required to balance the trade-offs between the use of open and close-ended activities to facilitate

learning while assessing student knowledge.

Part II

Crowdsourcing Instruction at Scale

Chapter 4

HOW COMMON ARE COMMON WRONG ANSWERS? CROWDSOURCING REMEDIATION AT SCALE

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

Proper citation for this chapter is as follows:

Gurung, A., Baral, S., Lee, M.P., Sales, A.C., Haim, A., Vanacore, K.P., McReynolds, A.A., Kreisberg, H., Heffernan, C., & Heffernan, N.T. (2023). How Common are Common Wrong Answers? Crowdsourcing Remediation at Scale. In Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23).

4.1 Introduction

The intricacies of learning mathematics are cognitively complex. Solving math problems demands students to understand the problem's requirements and demonstrate their knowledge and comprehension of the

topic [334]. Often, the problem-solving process involves breaking down the task into smaller sub-tasks that span several underlying concepts [327, 53]. This synthesis stage includes practicing various mathematical syntaxes, rules, and operations. The practice of synthesizing solutions reinforces students' knowledge and comprehension of the underlying concepts, thereby facilitating the development and consolidation of their understanding of mathematical principles [186, 337].

While the learning and synthesis processes may seem intuitive and straightforward, their analysis presents significant challenges [335]. The learner's individual problem-solving steps are intrinsic and can be challenging to deconstruct. Students can apply their inherent cognitive abilities to adopt different approaches towards solution synthesis [336, 74]. These approaches can vary, for example, in the complexity of the broken-down sub-task or the order in which the sub-tasks are solved [54].

Despite variations in approach, a fundamental understanding of mathematical processes is essential for problem-solving. However, gaps in knowledge, misconceptions, or "slips" can lead to incorrect responses [22]. Alternatively, insufficiently understood concepts may prompt students to guess answers or adopt incorrect problem-solving strategies, leading to a different set of errors [53]. Regardless of the cause, without directed feedback on how to resolve errors experienced during problem-solving, the errors may impede a student's learning progress. Understanding the common errors that students experience as they interact with mathematical problems is critical for guiding the design of effective instructional practices to help students learn correct mathematical processes and problem-solving strategies [257].

The diagnosis and examination of "Common Wrong Answers" (CWAs) is critical to understand learning processes in the context of mathematics. CWAs can be used to enhance educational technologies that, in conjunction with teachers, can address the needs of individual students—educational technologies often referenced as Computer-Based Learning Platform (CBLP), Online Learning Platforms (OLP), or Intelligent Tutoring Systems (ITS). For consistency, we will reference them as CBLP throughout this paper.

In a previous study, the authors of this paper examined the efficacy of two distinct types of Common Wrong Answer Feedback (CWAF)—verbose and detailed versus short and concise (c.f., [141]). The study employed a randomized control trial, where the control was business as usual, with no CWAF. The CWAs were proactively identified using a diagnostic model approach [53], and teachers, alongside learning activity designers, were tasked with generating the corresponding CWAFs. The analysis led to interesting insights for students working on mastery-based activities. The verbose and detailed feedback detailing both correct and incorrect steps undertaken by the students was detrimental to the student's likelihood of achieving mastery. On the other hand, short and concise CWAFs, while not significant, hinted towards a positive trend in

facilitating student mastery.

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

4.1.1 Research Questions

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

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

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

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

4.2 Background

4.2.1 Common Wrong Answers

Wrong answers are mistakes or errors that students typically make due to buggy rules, misconceptions about the topic, or gaps in knowledge. These CWAs have been the subject of substantial research in the fields of cognitive science and mathematical learning [54, 53, 394, 58, 387, 258].

Prior research [81, 329] has explored the correction of these common errors through instructional strategies. For instance, Brown et al., (1978) [53] analyzed frequent student errors when solving multi-digit subtraction problems and developed a diagnostic model that detects and elucidates these errors. Building on this,

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

Brown et al., (1980) [54] introduced the “generative theory of bugs,” a set of formal principles devised to explain the prevalent errors in procedural skills.

In their study, Sison et al., (1998) [341] proposed student modeling techniques to identify common errors in student work. They emphasized the need to assemble a “bug library,” a collection of the most common misconceptions or errors made by a specific student population. However, they acknowledged the challenges in creating these libraries, as misconceptions vary depending on the student population, and different student groups may demonstrate unique types of misconceptions when solving mathematical problems.

In addition to the principles of learning theory and cognitive skill acquisition, research has also investigated the potential of algorithmically identifying common student misconceptions to rectify incorrect and buggy processes in students’ work [329, 268]. Selent et al., (2014) [329] employed machine learning methods to predict CWAs and their underlying causes. They examined the effectiveness of providing *buggy messages* when a student makes a CWA. Their data suggested that these *buggy messages* led to a reduction in help-seeking behavior on a CBLP, indicating a possible rectification of common errors in students’ work.

4.2.2 Feedback Intervention

Feedback is a significant factor influencing learning outcomes and achievement. However, the impact of feedback is contingent on its type and mode of delivery. Previous research on Feedback Interventions (FI) through meta-analyses has produced mixed results regarding their effectiveness on student performance [217, 343, 318, 352, 28, 27, 192, 152]. These results have spurred further research to explore the intricacies of FI, culminating in the development of Feedback Intervention Theory (FIT) [192]. FIT posits that FIs aim to capture the recipient’s attention across three hierarchically organized levels: task learning, task motivation, and meta-task. While there are concerns about the general effectiveness of FIs [152], these concerns are less significant in an educational context as they have been found to be more beneficial in instructional settings. In a comprehensive synthesis of over 500 meta-analyses on the effects of schooling, Hattie (1999) (c.f., [152]) identified FIs as among the top 10 most influential factors on student achievement, thereby underscoring their effectiveness in promoting learning.

Effective feedback can help learners track their progress, validate their efforts, reinforce their progress, and impact their reactions and behavior when working on activities [398, 142, 67]. Feedback is indeed crucial to the student’s learning experience, but the quality of the feedback varies greatly. The effectiveness of feedback is often influenced by student perception. Some studies have reported on constructive feedback from instructors to be the most beneficial [376]. Conversely, if the feedback was too vague or lacked content,

its usefulness would diminish. Studies, such as [207], discuss how providing feedback in an online setting is an art and that there are various best practices including generating positive feedback and/or balanced feedback.

In this paper, we focus on the exploration of tailored feedback for the remediation of common errors, CWAs, in students' work. We adopt the Hattie et al. (2007) [153] conceptualization of feedback ², that expanded upon the generalized FIT model and proposed a theoretical model aiming to reduce the discrepancy between the current and desired understanding of learners in an educational context. Figure 4.1 presents the theoretical feedback model proposed by Hattie et al. [153] for enhancing learning. The model posits that the feedback must answer three major questions: (1) What are the goals? (2) What progress is being made toward the goal? (3) What activities need to be undertaken to make better progress?

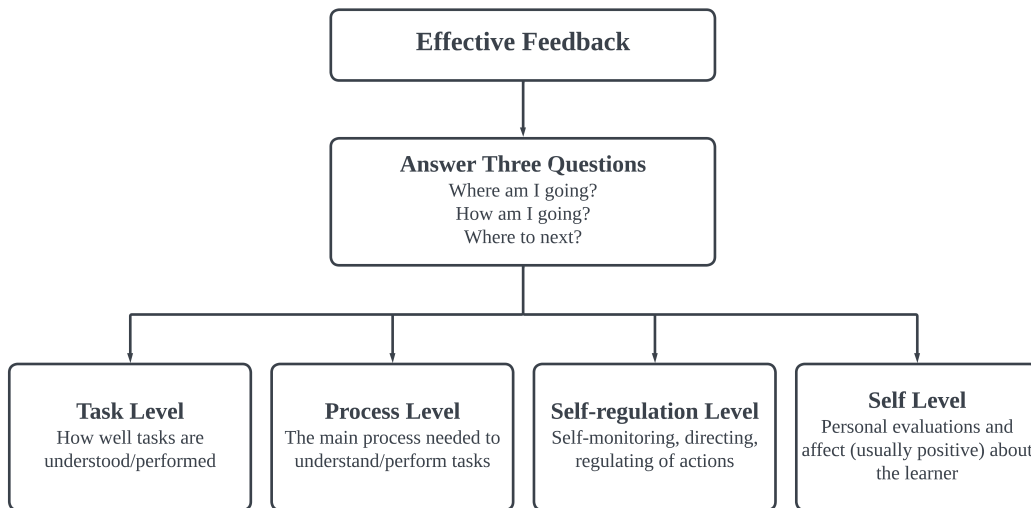


Figure 4.1: A model of feedback for enhanced learning, taken from Hattie et al. (2007) [153]

The FIs address these questions by operating across four levels of instruction: (a) task level, (b) process level, (c) self-regulation level, and (d) self-level. Therefore, effective feedback should recognize if the task requirement is understood, demonstrate the correct processes required to complete the task, include instructions that direct the learner towards the next productive actions, and include evaluation and affect (usually positive) to personalize the instruction.

²[153] Feedback is conceptualized as information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding. A teacher or parent can provide corrective information, a peer can provide an alternative strategy, a book can provide information to clarify ideas, a parent can provide encouragement, and a learner can look up the answer to evaluate the correctness of a response. Feedback thus is a "consequence" of performance.

4.2.3 *Common Wrong Answer Feedback*

Prior research has dedicated significant focus to the remediation of common errors in students' work [249, 250]. A study by Vanlehn et al. (2003) [362], for instance, evaluated the interplay between expert human tutors and physics students, specifically examining the efficacy of tutor explanations in rectifying student errors. The study reported that only certain explanations led to improved learning, with the effectiveness of feedback heavily contingent on the content and the question at hand. Moreover, shorter and more precise explanations were observed to be more effective than their longer, more elaborate counterparts. Thus reinforcing our prior work exploring CWAFs, where long and verbose CWAFs were detrimental to student mastery rates on mastery-based activities [141].

Additional studies have indicated the limitations of guided instructions in rectifying errors originating from misconceptions of previously learned skills [326]. These findings suggest that deeply ingrained misconceptions and errors might pose substantial difficulties to rectify over time.

Further research has proposed the use of error analysis methods as an essential step towards understanding students' ability to identify and explain errors in problems [198, 139, 319]. These studies involved presenting students with erroneous examples and requiring them to identify and articulate the errors within them. In particular, Rushton et al. (2018) [319] reported that this approach to error analysis led to better knowledge retention compared to traditional methods of learning mathematics.

4.2.4 *Crowdsourcing Instruction*

Crowdsourcing has emerged as a prevalent method in K-12 education for gathering feedback on instructional materials [103, 377, 187]. Leveraging various authoring tools, educators can create and disseminate educational content that is more representative. A variety of CBLPs and tools have integrated the crowdsourcing approach to encourage instruction and teacher-authored content [154, 39, 93, 274, 373, 384].

Research underscores the potential of crowdsourcing in enriching online learning experiences. It enables on-demand teacher support, tutoring, provision of hints, and explanations [289, 274, 241, 162, 384, 90]. Moreover, several studies have explored the use of crowdsourcing to collect teacher-given scores and feedback messages (instructive guidance) for students' answers on open-ended math problems to develop automated grading and feedback generation using Natural Language Processing (NLP) algorithms [47, 30]. The effectiveness of crowdsourcing in enhancing instructional materials and student learning experiences on online platforms has been well-documented [289, 274].

Building on these insights, our current study aims to crowdsource CWAFs by developing a platform for

teachers to identify and rectify CWAs.

4.3 Exploring Common Wrong Answers

To answer **RQ 1**, we explored the commonality of CWAs by examining data from students in grades 6, 7, and 8 who worked on problems in two commonly used curricula for mathematics in the US: Illustrative Mathematics (IM) and EngageNY (ENY) over a five-year period from ‘15-‘16 to ‘19-‘20. The students’ data were collected from ASSISTments [154] learning platform. A summary of the total number of problems the students worked on across the 5 school years from ‘15-‘16 to ‘19-‘20 is presented in table 4.1—the problems were considered eligible for the count if they were worked on by more than 20 students in at least one of the 5 school years. We observe that ENY on average is used more often than IM and on average teachers have used the content for grade 7 ENY the most across the 5 academic years.

Table 4.1: Summary of Total Problems and Problems with CWAs. The problems with CWAs met our threshold of more than 20 students working on the problem in two or more academic years.

<i>Academic Level</i>	Engage NY		Illustrative Math	
	Total Problems	Problems with CWAs	Total Problems	Problems with CWAs
Grade 6	1351	210	2082	254
Grade 7	1845	511	2088	518
Grade 8	1076	92	1475	267

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

Figure 4.2 provides an example from the EngageNY (ENY) curriculum, where two consecutive problems are associated with the same Common Core Standards, hence demanding a similar skill set. The first problem calls for the simplification of an equation, while the second entails verifying the results derived from the initial problem. Problems sharing a common skill set, like the ones mentioned, offer a greater likelihood of knowledge transfer compared to those derived from different Common Core Standards.

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

Problem 1 Standards 7.EE.A.2

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

$(5r)(-2)$

Type your answer below (mathematical expression):

Modified from EngageNY ©GreatMinds Full Attribution

Problem 2 Standards 7.EE.A.2

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

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

Modified from EngageNY ©GreatMinds Full Attribution

Figure 4.2: An example of two consecutive problems from ENY Grade 7 Module 3 Lesson 1 where both problems have the same set of Common Core Standards.

In our analysis, we found that 1,045 problems had CWAs spanning at least two academic years. Table 4.2 provides an example of these CWAs across academic years for the second problem presented in figure 4.2, from ENY grade 7 module 3 lesson 1. As reported in table 4.2, we observe that the first CWA met the commonality threshold in four out of the 5 academic years, indicating consistency. However, the second and third CWAs demonstrated some fluctuation, with ranks interchanging in some years, and entirely new CWAs appearing in others.

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

Our exploratory analysis of the occurrence of CWAs revealed a pattern of repetition across academic years. A more in-depth analysis of the problems featuring CWAs indicated that the majority of the problems belonged to “*Practice Problems*” (in IM) and “*Problem Sets*” (in ENY)³. As the term problem set is

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

Table 4.2: Common Wrong Answer by Student Count on the second problem as presented in figure 4.2. The threshold for the CWA requirement was met in 4 of the 5 academic years from ‘15-‘20. The threshold required more than 20 students to work on the problem in each academic year with more than 10 students making the same CWA.

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

generally used to represent a set of problems that can be assigned to students, we will refer to both *Practice Problems* and *Problem Sets* activities as *Practice Problems* throughout this paper.

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

4.4 Task Abstraction

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

During the goals analysis, we broke down each goal into a series of sub-goals that directly align with teacher needs. For instance, a high-level goal might be: facilitate effective feedback, which could be broken down into sub-goals such as ‘analyze student error rates’ and ‘allow teachers to easily input their feedback’. We utilize the sub-goals to identify the visualization components needed in the crowdsourcing tool to meet teachers’ needs effectively. We utilized the “Nested Model for Visualization” (c.f., [246]), a common Human-Computer Interaction (HCI) technique, to identify the fundamental goals of a crowdsourcing tool.

Upon validating the high-level goals and sub-goals with end-users and domain experts, we proceeded with task analysis, defining low-level tasks allowing browsing, exploring, and identifying various aspects of the data to facilitate the sub-goals. These tasks, derived from the Brehmer and Munzner topology (c.f., [48]),

provided a useful roadmap for designers and developers during the tool's creation. While our crowdsourcing tool doesn't include the elaborate visualization components associated with common HCI projects, the Nested Model for Visualization, and Brehmer and Munzner's topology proved invaluable in identifying the tool's fundamental goals and tasks, which ultimately helped in enhancing teachers' ability to formulate effective feedback.

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

4.4.1 Goal Analysis

Table B.1 lists the goals and sub-goals resulting from our analysis. The overarching goal of the tool is to augment teacher ability in gaining insight into the various processes the students might have taken during the synthesis of a solution that resulted in the CWAs. While the underlying mechanism that resulted in the CWAs is unknown, we aim to leverage teacher experience and intuition to discern the underlying cause and generate appropriate feedback to help remedy the cause.

We identified 3 distinct goals a crowdsourcing tool needs to facilitate. The first two goals, G1, and G2, directly address teacher needs in substantiating the CWAs and providing contextual insight to help teachers formulate effective feedback. Goal 1 helps teachers understand the general student performance on the problem, provide evidence towards the commonality of the response, and identify the problems within a set of problems where students struggle the most, i.e., most likely problems within a set of problems where gaps in student knowledge will impact their performance the most.

The intent of goal 2 is to provide contextual information that can augment teacher ability when analyzing the CWAs and their potential causes by providing contextual information. Additionally, information on prior problems related to the same skill component can provide scaffolding that teachers can leverage in contextualizing the problems and converging on a smaller subset of potential causes for the CWAs.

While the primary objective of the tool is to facilitate the generation of CWAFs, both the teachers and domain experts on multiple occasions throughout the task abstraction processes emphasized the importance of goal 3 in fostering self-actualization for teachers through collaborative feedback enhancement. It enriches their participation in a generation of CWAFs through peer support and fostering a sense of camaraderie. Such opportunities allows teachers to contribute to and benefit from the collective knowledge.

Table 4.3: Fundamental goals of a crowdsourcing tool.

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

4.4.2 Task Analysis

For each sub-goal presented in table B.1 we generated a list of low-level sub-tasks designed to help teachers (a) look up other problems within the problem set, (b) explore various knowledge components the students struggled with while working on the problems, (c) identify the potential causes of the CWAs, and (d) produce feedback that can effectively help remediate gaps in student knowledge that resulted in the CWAs. These sub-tasks are related to the abstract visualization task from Brehmer, and Munzner's topology [48].

Table B.2 illustrates high-level tasks that can guide the design and development of features in the crowdsourcing tool, facilitating one or more sub-goals. Together, these tasks contribute to achieving the main goals of the crowdsourcing project. While these tasks can be further decomposed into more specific sub-tasks, we focus only on high-level tasks to avoid unnecessary complexity. We believe these tasks are self-explanatory and refrain from extensive elaboration to conserve space and prevent redundancy.

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

Table 4.4: Task analysis deconstructing the feature requirements of each sub-goal.

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

4.5 Crowdsourcing Common Wrong Answer Feedback

In this section, we briefly describe our implementation of the crowdsourcing tool guided by the goals and task analysis described in the prior section. In order to facilitate the fundamental goals described in table B.1 we designed a new crowdsourcing platform within the ASSISTments ecosystem. The tool allows teachers to identify relevant CWAs, gain contextual insight into the problems associated with the CWAs, and facilitates peer collaboration to help further improve the quality of the CWAs.

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

PRABEUZ2 - Part1
 Problem Type: Number

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

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

Based on these results, estimate the probability that 3 or more kicks are made.
 copied for free from openupresources.org

CWA-feedback.
 Number of students who responded: 273 Percent Correct: 50.54%
 Answer: 4/5

First Common Wrong Answer feedback for students

Common Wrong Answer: 4
 Number of Students: 50/135
 Percent of Students: 37% of incorrect responses were this answer.
 Teacher Name wrote:
 Good job noticing that the number of desired outcomes is 4. Express the desired number of outcomes as a fraction of the total possible outcomes

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

Reviewer/Commenter Name

Second Common Wrong Answer feedback for students

Third Common Wrong Answer feedback for students

Figure 4.3: Teacher perspective, visualization of a problem from Illustrative Math curricula with Common Core standard 7.SP.C.8.b where a teacher has written feedback and a peer/moderator has reviewed it as well.

Figure 4.3 displays the teacher perspective on a problem set in IM curricula for grade 7, unit 8, lesson 8–based on the common core standard for “Probability and Sampling”. As the figure illustrates a teacher has analyzed the first CWA for the problem and provided appropriate CWF. The teacher can substantiate the CWAs, **Goal 1**, by examining the number of students that have worked on the problem, the percentage of students who answered it incorrectly, identifying the top 3 CWAs, and the percentage of students who made the CWAs among students who answered it incorrectly.

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

The primary focus of this paper is to analyze CWAs and evaluate the efficacy of CWFs in addressing the underlying causes of the CWAs. We collaborated with 24 experienced middle school teachers using IM or ENY in their classrooms. These teachers were tasked with generating CWFs for Grade 7 *Practice Problems*. To ensure the feedback aligned with the curriculum requirements, teachers received preliminary

training from domain experts. The experts also offered continuous feedback and served as moderators during the crowdsourcing process to maintain the quality of CWAFs. After the CWAFs were crowdsourced, the experts performed a final review to approve the feedback, marking it as ready for student use.

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

4.6 Implementing Common Wrong Answer Feedback

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

4.6.1 *Experimental Design*

Once the students start a problem, students are randomized into either a control group, business-as-usual (no CWAF), or a treatment group (receiving CWAFs). Ideally, randomizing students once they make a CWA would be optimal; however, the process of triggering a server request that randomizes students once they enter a CWA can take away from the learning experience of the student and can ultimately hamper their perception and usage of the platform itself as such we randomize beforehand and analyze the effectiveness of CWAFs on the treated group. We implemented a 90:10 randomization split, providing a 90% chance of a student being assigned to treatment and a 10% chance to control. This ratio was strategically chosen to optimize access to learning opportunities for as many students as possible.

4.6.2 *Dataset*

Since the initial implementation of the first batch in April '22, CWAFs have been randomized across 20,044 students working on 1,387 problems in ENY and IM a total of 623,857 times; students were assigned 560,897 times to treatment and 62,960 times to control. While the students were assigned to treatment or control, they only received CWAFs if their attempt was one of the top 3 CWAs for the problem. As such, we dropped the students who did not attempt to answer the problem with a CWA at any point while working on the problem. After dropping the students who did not make any attempts that identified as a CWA for both control and treatment, we have 14,672 unique students who were randomized and made at least one CWA when working across 947 problems. With this, we have 96,398 instances of students randomized to treatment

and 10,960 to control. As we used a 90:10 randomization design, we explored the balance across conditions by conducting a binomial hypothesis test on the next problem attempt after receiving a CWAF. Our sample failed the binomial hypothesis test indicating an imbalance across the attrition rates for treatment and control, as such we scored 0s for instances where the students dropped out without attempting the next problem. While this data is for students working on problems within the same problem set, different problems within a single problem set can have different sets of common core standards. As such, we filter the treated students to examine the effectiveness of CWAFs by only analyzing the problems where both the intervention and the next problem had the same common core standards. This additional filtering requirement reduced the number of distinct students to 12,175 and the number of distinct problems to 535, where students were randomized 62,688 times into treatment and 7,080 times into control.

4.6.3 Evaluating the Effectiveness of Common Wrong Answer Feedback

For answering **RQ 3**, in this section we analyze the efficacy of CWAFs in the remediation of common wrong answers (CWAs). We explore this by examining the binary correctness of the next problem using the *lme4* package in R. We use a pre-registered logistic regression model to explore the effectiveness of CWAFs ⁴. The pre-registered logistic regression model is listed in equation 8.1.

$$\begin{aligned} \text{next problem correctness} \sim & \text{treatment} * \text{prior 5 problem avg correctness} \\ & + (1|\text{CWA writer}) + (1|\text{problem}) + (1|\text{class}) \end{aligned} \tag{4.1}$$

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

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

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

the impact of classroom-level factors, as students' motivation and learning behaviors are often influenced by their relative standing among their classmates.

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

4.6.3.1 *Between Consecutive Problems with the same set of Common Core Standards*

For the problems within the same set of common core standards within the consecutive problems (within-skill), the results from the regression analysis are reported in table 4.5. We observe that students in the treatment condition had significantly higher odds to answer the next problem correctly for the problems with the same set of common core standard tags (Odds-Ratio = 1.07, p-value = 0.028). The fixed effect of mean-centered prior 5 problem average correctness was significant and highly predictive of next-problem-correctness. While CWAFs do appear to have a net positive benefit, there was a significant interaction between treatment and prior 5 problem average correctness indicating a potential heterogeneous treatment effect ⁵.

4.6.3.2 *Between Consecutive Problems in the same Assignment irrespective of Common Core Standards*

For the problems irrespective of the common core standards within the consecutive problems (within-assignment), the results from the regression analysis are reported in table 4.6. We observed similar results on the other covariates; however, while leaning in the positive direction we did not observe a significant difference between students in control and treatment, indicating that the transfer of knowledge in consecutive problems to be inconclusive (Odds-Ratio = 1.03, p-value = 0.188). The fixed effect of mean-centered prior 5 problem average correctness was significant and highly predictive of next-problem-correctness, however the interaction between treatment and prior 5 problem average correctness while similar to the within-skill analysis was not significant ⁶.

4.7 Discussion and Future works

Our analysis revealed a relative consistency in the incorrect answers made by students across academic years. While the same CWAs were not the most common for the same problems in every school year, there

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

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

Table 4.5: Exploring the effectiveness of CWAf by using next-problem-correctness(binary) as a dependent measure for the same set of Common Core Standards (within-skill) in consecutive problems.

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

Table 4.6: Exploring the effectiveness of CWAf by using next-problem-correctness(binary) as a dependent measure within-assignment irrespective of the set of Common Core Standards associated with consecutive problems.

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

was an obvious pattern indicating an overlap in the top 3 CWAs. We also observed that teachers using IM and ENY prefer to assign *Practice Problems* over *Exit Tickets*, *Student Facing Task*, and *Cool Down* problem sets. While various prior works exploring CWAs in the past have expressed concerns regarding the reliability of CWAs [53, 359], our analysis substantiates the commonality of CWAs. A potential cause of the replication challenges encountered by prior works [362] exploring the reliability of CWAs could be attributed to the smaller sample size, as our analysis does indicate the prevalence of CWAs at scale. It is important to note that our work does not claim to provide insight into the various underlying mechanisms students utilize when synthesizing solutions that can result in the incorrect answer due to “bugs” in their processes, but rather through this work, we aim to establish the reliability of the CWAs that can be caused by gaps in student knowledge, misconceptions, guess, slip, or bugs when formulating solutions.

While the primary objective of this paper was to explore the fidelity of CWAFs, in this paper, we also wanted to focus on various design and development techniques that can be potentially beneficial to future research. While the Learning@Scale (L@S) community at large has designed and successfully developed systems at scale, it is noteworthy that there has been a limited emphasis within our community on documenting the various design and development principles that inform the successful implementation of such systems. As such, in this paper, we leverage the design philosophy commonly used in visualization projects to conduct task abstraction that can elucidate the various aspects of crowdsourcing that are fundamental in the overall successful adoption of such tools. In our case, the objective was to develop a tool that can augment teacher ability to examine CWAs when writing CWAFs. The primary benefit of the goals and task analysis is to identify critical features a tool should facilitate and the hierarchy of such features to ensure the successful implementation of the tool. As such, this paper presents the fundamental goals and tasks a crowdsourcing tool needs to facilitate a successful adoption. Each goal is designed to build on prior goals and further enhance the process of facilitating crowdsourcing. While there is no evidence to suggest that the design philosophy used in the development of this crowdsourcing tool led to the creation of more effective feedback in comparison to other design philosophies, we did observe that the CWAFs lead to positive learning outcomes across consecutive problems focusing on the same skill set. This positive outcome is particularly important in the domain of CWAFs research as there is mixed evidence regarding the fidelity of CWAFs, with some reporting positive results [249, 250, 362]. In contrast, others have reported on the lack of benefit in using CWAFs [326, 141]. A well-designed system can provide powerful affordance that can enhance the quality of the outcome by facilitating exploration, learning, and collaboration when leveraging crowdsourcing.

As attested by the lack of variance in the outcome due to CWA writer, as random intercepts, in both the

within-skill and within-assignment models reported in table 4.5 and table 4.6 respectively. This observation suggests that the training and use of moderators to generate a consistent set of CWAFs, following the principles outlined by Hattie et al. (2007) [153] as presented in figure 4.1, was successful. In future work, we intend to leverage the CWAFs generated through moderated crowdsourcing as a baseline when comparing the effectiveness of different CWAF designs. As these CWAFs were generated across 1,660 problems, we can now hypothesize and test the effectiveness of different types of feedback across different topics and subfields of mathematics, e.g., geometry, statistics, algebra, and arithmetic.

In our final analysis, we examine the effectiveness of CWAFs by examining the transfer of knowledge on the next problem using the binary measure of the next-problem-correctness in two contexts, within-skill, and within-assignment. Our findings reveal that students appear to benefit from CWAFs, as evidenced by their increased likelihood of solving consecutive problems correctly within-skill. This outcome is noteworthy, particularly in the context of IM and ENY curricula, where subsequent problems within a skill set tend to increase in difficulty. However, we did not observe a similar benefit on subsequent problems within-assignment. These findings suggest a contextual aspect of the effectiveness of CWAFs. Further investigation is needed to develop our understanding of these dynamics. For instance, while the within-skill knowledge transfer could occur due to the CWAFs effectively addressing student needs, it is also entirely plausible that the CWAFs are causing shallow learning—as evidenced by the lack of knowledge transfer within-assignments. Additionally, further analysis exploring learner behavior around CWAFs is required to understand if students are attentive to the CWAFs. A prior analysis has explored student attention towards hints by utilizing response time decomposition, where higher attention to hints was correlated with student learning outcomes [142].

While the focus of this paper has been the exploration of CWA and the efficacy of crowdsourced feedback, we implore fellow researchers and developers in our L@S community to consider leveraging similar task abstraction methodologies in their own work. We believe the insights provided in our goal analysis, presented in Table B.1, can serve as initial guardrails for informing future research aimed at developing tools exploring similar crowdsourcing challenges. Such methodologies can potentially streamline the process of identifying the fundamental features in crowdsourcing contexts, thus enhancing overall efficiency and output.

4.8 Conclusion

At the onset of this research, we posited the existence and prevalence of CWAs in a learning context. Our findings substantiate our initial hypothesis, revealing a remarkable persistence of CWAs across different academic years, even with changing student populations. Utilizing this understanding, we successfully devel-

oped a new crowdsourcing tool to facilitate the collection of Common Wrong Answer Feedbacks (CWAFs) from educators. Our analysis demonstrates that the integration of these teacher-generated CWAFs leads to improved learning outcomes, particularly evidenced by the observed transfer of knowledge across consecutive problems that focus on the same skill set (within-skill). Interestingly, the effectiveness of CWAFs was less pronounced when consecutive problems irrespective of associated skill sets (within-assignment). This distinction offers a promising avenue for further investigation in future studies. Furthermore, our work has produced a baseline that can be leveraged by future research exploring CWAFs.

Part III

Use of Automated Grading and Feedback Generation on Open Response Problems in Mathematics

Chapter 5

INVESTIGATING PATTERNS OF TONE AND SENTIMENT IN TEACHER WRITTEN FEEDBACK MESSAGES

Feedback is a crucial factor of student learning in mathematics. Whether in the form of simple indicators of correctness or textual comments, feedback can help guide students' understanding of content. Beyond this, however, teacher-written messages and comments can provide motivational and affective benefits for students. With this, the question emerges as to what constitutes effective feedback to promote not only student learning but also motivation and engagement. Teachers may have different perceptions of what constitutes effective feedback and utilize different tones in their writing to communicate their sentiments. This study aims to investigate trends in teacher sentiment and tone when providing feedback to students in a middle school mathematics class context. Toward this, we examine the applicability of state-of-the-art sentiment analysis methods in a mathematics context and compare correlations of this measure with student performance metrics. In addition, we explore the use of punctuation used in teacher feedback messages as a measure of tone. Finally, considering the subtle conceptual differences between tone and sentiment, we examine whether our measures of these constructs correlate with each other as well as other aspects of the feedback message.

Proper citation for this chapter is as follows:

Baral, S., Botelho, A.F., Santhanam, A., Gurung, A., Erickson, J., & Heffernan, N.T. (2023). Investigating Patterns of Tone and Sentiment in Teacher Written Feedback Messages. In The 24th International Conference on Artificial Intelligence in Education (AIED 2023).

5.1 Introduction

Feedback is an essential part of student learning. Whether in the form of simple indicators of correctness or more descriptive textual comments, feedback can help guide students' understanding of instructional content, offer solutions to fix errors in their work, and provide motivational and affective/emotional benefits to the students, improving their overall learning experience. Some teachers may prefer to use a more directive approach when giving feedback, while others may take a more supportive approach. Additionally, the approach used by teachers may differ based on different groups of students, such as the students who are struggling versus those who are exceeding in their given task.

In designing tools to support the provision of feedback for teachers in the context of online learning platforms, it is important to understand not only how to structure feedback so that it is effective in improving student learning, but that feedback also needs to match the teacher's voice so that they want to utilize it. Teachers may have different communication styles, and they tailor their approach of feedback to meet the needs of their students. Toward this, understanding the sentiment and tone carried by teachers' feedback to students is necessary. While prior works have examined the analysis of sentiment in various domains (e.g. [117]), this work observes a subtle distinction between this concept and that of tone. While sentiment refers to the emotional valence of the text itself, we define tone as the intended emotional response to the feedback. Consider, for example, a teacher who provides the feedback of "Come on, I know you can do this!" to a student who responded to a problem with an answer such as "I don't know". While, without context, the sentiment of the text itself is arguably positive, in reality, the tone is more critical in nature.

The study aims to investigate trends in teacher-written feedback messages in a middle school mathematics context through sentiment and tone analysis of these comments. Through examination of the applicability of state-of-the-art sentiment analysis methods and exploration of the use of punctuation in teacher feedback messages, this study aims to gain a deeper understanding of how teachers choose to structure their feedback. Additionally, by considering the subtle conceptual differences between tone and sentiment, we explore whether our measures of these constructs correlate with each other and other aspects of the feedback message or student performance metrics. By examining these trends, we hope to gain a better understanding of the impact of feedback on student learning in mathematics and inform recommendations for best practices in the delivery of feedback. As such, our main research questions are:

1. How well do the state-of-art sentiment analysis methods perform when predicting sentiment in a mathematical context?
2. Which punctuation marks are frequently used in teacher feedback messages and how do they relate to measures of students' performance when used as a measure of tone?
3. How do sentiment and the use of punctuation marks in feedback messages correlate with each other and other aspects of feedback and students' performance?

5.2 Background

Researchers in the past have reported on meta-analyses exploring the effects of Feedback Interventions (FI) on performance, with mixed results suggesting that the context, content, and structure of feedback impact

its effectiveness [152, 192, 343, 217, 318, 28, 27, 352]. Such inconsistencies have led to further research exploring the nuances of FI that resulted in the development of Feedback Intervention Theory (FIT; [192]). FIT operates under the assumption that FIs aim to catch the recipient's attention across 3 hierarchically organized levels: task learning, task motivation, and meta-task. While there are concerns regarding the general effectiveness of FIs it is much less of a concern in an educational context. Hattie [152] reported on a synthesis of over 500 meta-analyses¹ exploring the effect of schooling on students where meta-analyses exploring the effectiveness of FIs [343, 217, 318, 28, 27, 352] found them to be among the top 10 highest influences on the student achievement—highlighting the effectiveness of FIs in learning.

Feedback can often impact students' reactions and behavior when working on activities[398, 142, 67]. Student perception plays a crucial role in the effectiveness of the feedback; as reported by Weaver and colleagues [376], students who perceived feedback as vague or lacking content exhibited little benefit as compared to students who recognized feedback as detailed and constructive. Studies, such as [207], discuss that providing feedback in an online setting is an art and that there are various best practices, including generating positive and/or balanced feedback (positive, negative, then positive). Hattie et. al [153] posit that effective feedback must answer three major questions: 1) What are the goals? 2) What progress is being made toward the goal? 3) What activities need to be undertaken to make better progress? Effective feedback must incorporate features such as recognize if the task requirement is understood, exhibit the correct processes required to complete the task, include instructions that direct the learner toward productive action.

Growth and innovations in the field of Educational Technology (Ed-Tech) have influenced the adaptation and regular usage of Computer Based Learning Platforms (CBLPs) in classrooms. Research on CBLPs has focused on automating the scoring of open response problems that require students to provide verbose responses that are semantically structured[188, 7, 57, 306, 397]. Some researchers have attempted to automate the scoring of open-response problems in mathematics. Studies such as [199] attempted to automatically score student open-response answers in mathematics; however, this study removed any non-mathematical content. Similarly, others have explored the implementation of Natural Language Processing (NLP) on open-response problems in mathematics, including non-mathematical text [113, 30]; however, the focus has primarily been on automating the scoring and not the feedback generation processes. Open-ended responses in mathematics can drastically differ from those in non-mathematical domains as open-ended essays. Short answers in non-mathematical domains often comprise multiple sentences and paragraphs[68, 97, 306, 397], whereas responses in mathematics generally are more concise and often incomplete sentences[199, 113]. De-

¹The meta-analysis by Kluger et al, 1996[192] that proposed FIT was included in the synthesis meta-analysis

spite the sparse responses, teachers intuitively infer the students’ understanding of the topic and deconstruct their approach to solving the problem—enabling them to formulate effective feedback. While there is no denying teachers’ ability to formulate feedback, automating feedback for open-response problems, especially in mathematics has presented a substantial challenge.

In this work, we leverage historical data on feedback provided to students working on middle school math problems on a CBLP. We primarily focus on understanding the various problem, student, and teacher-level factors that influence the sentiment and tone of the feedback. In the following sections, we explain in detail the dataset used, then the analyses to examine the patterns in sentiment and tone of teacher-provided feedback messages.

5.3 Dataset

The study uses a teacher feedback dataset taken from ASSISTments[154], consisting of student answers to open-ended math problems and teacher-authored textual feedback messages. The data includes 8,307 open-ended mathematics problems and 1,93,187 total responses given by 23,853 distinct students and the corresponding feedback message given by 1,296 different teachers. Data cleaning was performed to drop any data with empty feedback messages. The dataset consists of scores on a 5-point integer scale ranging from 0 to 4 provided by teachers through a manual scoring process as part of normal classroom instructional practices. Scores beyond this range were considered outliers and were dropped as a part of the data-cleaning step, and non-integer scores were rounded down to integer format (this affected less than 5% of data samples).

In addition to the assessment data, the dataset also contains measures of students’ prior knowledge (a measure of the average correctness score of students across all the prior problems they have solved within ASSISTments[154]) and whether or not a student completed the assignment. Further for analysis purposes, we drop all the students who have completed fewer than 5 problems within the platform. The resulting dataset consists of 1,86,073 feedback from 1210 teachers given to 22,022 different students on their work to 8,237 different open response problems.

5.4 Analysis 1: Sentiment Analysis in Mathematics

Toward understanding the sentiment of teacher-written feedback messages in mathematics, we conduct a sentiment analysis to infer whether a given feedback is ‘Positive’, ‘Negative’, or ‘Neutral’ using a fine-tuned downstream version of the ‘*bert-base-uncased*’ model [111]. This is a transformer-based model trained over a generic dataset of classified text. As most of the commonly-used sentiment analysis methods are based on

Table 5.1: Most common mathematical words picked from a list of the top 100 most frequent words in the teacher feedback messages dataset categorized by their sentiment.

Sentiment	Mathematical Words
Positive	value, side, multiply, explanation, ratio, equal, enter, label, length, solve, congruent, scale
Neutral	answer, number, line, point, +, -, equation, explain, angle, graph, question, divide, rotate, unit, slope, degree, reflect, factor, area, solution, first, segment
Negative	triangle, mean, reason, measure, problem

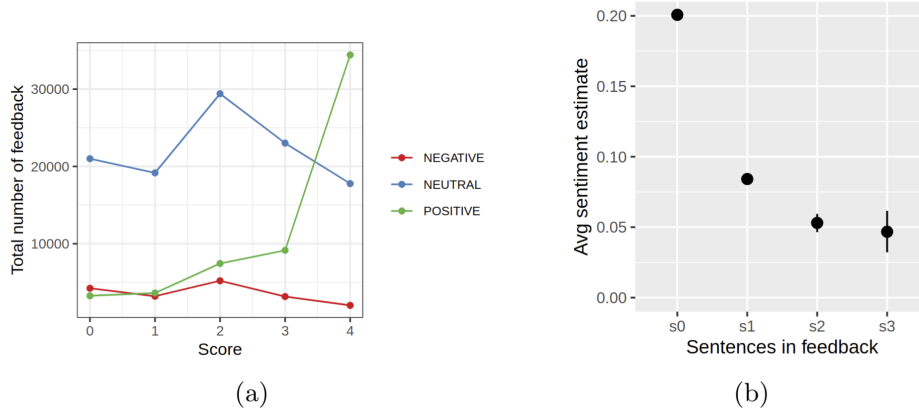


Figure 5.1: Plot showing the total number of feedback messages per score category grouped by the sentiment of feedback messages (a) and the average sentiment estimate across the first, second, third, and fourth sentences in the feedback message using continuous-valued outcomes from the model (b).

social media data, we hypothesized that this model being trained on a generic dataset had a higher likelihood of generalizing to our application domain (a hypothesis that will be tested).

We first seek to validate the use of a pre-trained sentiment model for use on our dataset by examining the impact that mathematical terminology may have on model estimates. A potential shortcoming of automated sentiment analysis methods is that such models may be confused by domain-specific language; this poses a potential risk in misinterpreting results. For example, words such as “power”, “addition”, and “multiply” may be associated with positive valence in certain contexts, but likely represent neutral mathematics concepts when used in the context of teachers’ feedback messages.

Considering the potential effect of some of these mathematical terms on the sentiment, in our next step we remove these common math words before predicting the sentiment of the feedback messages. For this,

Table 5.2: Some examples of Positive and Negative feedback messages from teachers, their sentiments with and without math terms, and their score.

Teacher-written Feedback	Sentiment w Math	Sentiment w/o Math	Score
[REDACTED] - you were doing a great job. Please don't enter nonsense responses.	Positive	Positive	0
I like that you labeled your angles with 3 letters. Angle CDM is 90 degrees. Angle DMC is 63 degrees. Together they make 153 degrees. Remember that complementary refers to 2 angles whose sum is 90. Can you find 2 angles that would add up to 90?	Positive	Positive	2
congruent	Positive	Neutral	3
Labels!	Positive	Neutral	4
Perfect Answer!!	Positive	Positive	4
-2; lack of effort in completing cool down.	Negative	Negative	0
This will cost you 2 points for Unit 5, lesson 8.	Negative	Negative	0
No - x would have to be negative.	Negative	Neutral	2
When we ignore the 5 or 6, we reduce the number outcomes down to 4 instead of 6. That way $P(\text{score})=1/4$ and $P(\text{not score})=3/4$.	Negative	Neutral	2
Label your units please	Negative	Negative	3
Sorry this was not working for you!	Negative	Negative	4

we first identify the top 100 most-frequent words from all the teacher feedback dataset, and from this list, we extract only the mathematical terms. Table 5.1 lists the common math terms extracted as a part of this step and categorizes them based on their predicted sentiment from the pre-trained model. We stem each of the extracted words to their base form (eg. multiply, multiplied, etc would be stemmed to multipli) and then exclude these terms from the feedback before finally applying the sentiment prediction model. Table 5.2 presents some examples of teacher feedback messages and their resulting sentiment with and without the

mathematical words. The sentiment distribution across various score categories is shown in Figure 5.1(a), and the average sentiment estimate across the first 4 sentences of a feedback message is presented in Figure 5.1(b), using continuous-valued outcomes from the model. From this, the first sentence is typically positive.

5.5 Analysis 2: Analyzing Tone using Punctuation Marks

Table 5.3: Percent of feedback with commonly used punctuation marks across each of the score categories.

Score	?	!	:) :-)
0	11.61%	3.36%	0.29%
1	16.38%	2.59%	0.54%
2	18.43%	3.48%	0.45%
3	17.19%	5.69%	1.10%
4	2.97%	41.12%	4.55%

The use of punctuation marks within a text of writing can reveal important cues about the tone and sentiment expressed in the text. For example, exclamation ‘!’ marks are used within a piece of writing to indicate the writer’s excitement, happiness, and sometimes, conversely, anger. Use of question ‘?’ marks, in the direct sense, indicate a question, but can also be a rhetorical approach to inspire thought or convey discontent (e.g. “???”).

For this Analysis, we look into the top 5 commonly used punctuation marks in the feedback messages which are: ‘.’, ‘;’, ‘?’, ‘!’ and ‘)’ respectively. From these, we focus on the use of question marks and exclamation marks. Also, we understand that the use of ‘)’, may be used by some teachers to express a smiling emotion, and in some other cases may be used in the form of mathematical expression. Question marks and exclamation marks are seen in about 12% and 15% of the feedback data respectively. Table 5.3 shows the use of some of these common punctuation marks across the feedback messages.

Continuing this analysis, we seek to identify first whether the usage of exclamation and question marks is most explained by student-, problem-, or teacher-related factors. In other words, we want to identify if the usage is more related to a communication style of the teacher or if instead factors such as problem difficulty or student ability explain their usage.

First, we perform a multilevel null model of logistic regression where the presence/absence of these punctuation marks as the dependent variable (in separate regressions), using no variables at level 1, and

teacher-, problem- and student-identifiers at level 2; this model will reveal which factor explains the majority of variance in the dependent. In the next step, we perform a multi-level logistic regression model with the presence of punctuation marks as the dependent variable and using different features of student and feedback such as prior knowledge, score on the current problem, assignment completion, and total words used in the feedback itself at level 1, and teacher-identifier at level 2. Exclamation marks, as are used both to express positive emotions like happiness and sometimes negative emotions like anger, we hypothesize that teachers use them differently based on the correctness (score) of the student in the given problem. Thus, we perform two separate analyses for low and high score categories – scores of 0,1,2 as low scores and scores of 3,4 as high score category.

5.5.1 Results of Analysis 2

5.5.1.1 Question Marks

Table 5.4: The resulting model coefficients for the multi-level logistic regression model on the use of question marks

	Variance	Std. Dev.
<i>Random Effects</i>		
Teacher	2.496	1.58
	β	Std. Error
<i>Fixed Effects</i>		
Intercept	-2.614***	0.082
Score	-0.718***	0.018
Total Words	0.195***	0.020
Prior Knowledge	-0.147**	0.052
Assignment Completion	0.235***	0.042

*p < 0.05 **p < 0.01 ***p < 0.001; β = standardized coefficient

The null model of multi-level logistic regression on the presence of question marks explains 63% (0.63 theoretical r-squared error) of the total variation in the data, indicating that higher level factors of teacher, problem, and students are a good predictor for the presence of a question mark in the feedback. With this however, most of the variance is explained by the teacher level (Variance = 3.18), then by the problem level (Variance = 2.13), and some are explained by the student-level identifiers (Variance = 0.28). This suggests that some teachers are likely to ask more questions than others, but this also depends on the problem.

The results for the multilevel logistic regression on the presence of question marks are presented in Table 5.4. The marginal r-square for this model is 0.02, suggesting that 2% variance is explained by the fixed effects variable, and the conditional variance is 0.44 suggesting that 44% variance is explained by the overall model including the level 2 random effects coming from teacher level identifiers. All the independent variables are significant predictors of the presence or absence of question marks in the feedback. Students with a high score on the current open-response problem and who have high prior knowledge, are less likely to get questions in the feedback, whereas students with a lower score and low prior knowledge are more likely to get questions in the feedback from teachers. Also longer the feedback with more words, the higher chance of the presence of question marks on them.

5.5.1.2 Exclamation Marks

Table 5.5: The resulting model coefficients for the logistic regression model on the use of exclamation marks categorized according to low and high score categories.

	Low Score		High Score	
	Variance	Std. Dev.	Variance	Std. Dev.
<i>Random Effects</i>				
Teacher	7.344	2.71	4.387	2.095
	β	Std. Error	β	Std. Error
<i>Fixed Effects</i>				
Intercept	-4.552***	0.218	-9.337***	0.161
Score	0.032	0.025	2.101***	0.031
Total words	-0.779***	0.054	-1.486***	0.026
Prior knowledge	0.020	0.012	0.230***	0.072
Assignment Completion	-0.184*	0.077	0.408***	0.049

*p < 0.05 **p < 0.01 ***p < 0.001; β denote standardized coefficient

For the null model of multi-level logistic regression for the use of exclamation marks, we see similar results as for question marks. This model explains 75% of the total variance in the dataset, with more variance coming from the teacher level (Variance = 6.23), then from problems (Variance = 1.53), and some from the student level (Variance = 0.57). For the multi-level logistic regression model, we have two separate

analyses for low and high-score categories of answer as presented in Table 5.5. For the low-score category, the marginal r-squared is only 0.009 suggesting that 0.9% variance is explained by the fixed effects of student and feedback level features, with the conditional r-squared error of 0.69 suggesting that overall 69% variance is explained by the model including the random effects of teacher identifier. Among the low-score category only, total words and assignment completion are the features found to be statistically significant. With more words used in the feedback, less chance of seeing an exclamation mark on them and vice versa. Similarly, the students who do not complete their assignments are more likely to see exclamation marks on their feedback. In the low score category, the score (either they get 0,1 or 2) and prior knowledge of students are not found to be statistically significant.

For the high score category, the marginal r-squared is 0.23 suggesting that 23% variance is explained by the fixed effects, and the conditional r-squared error of 0.67 suggesting that overall 67% variance is explained by the model including the random effects from the teacher level. All the fixed effects variables are considered to be statistically significant. Score on the current problem, prior knowledge of the student, and the assignment completion are all positively correlated with the presence of exclamation marks in the high score category. However, total words in the feedback is negatively correlated to the presence of exclamation marks, meaning exclamation marks are seen more on shorter feedback messages than in longer ones. The result also suggests that teachers are more likely to use exclamation marks for students who get a score of 4 than for students who get a score of 3. This is on par with what we have seen in most of the examples, where teachers use shorter feedback messages like 'Good Job!', and 'Great!', etc expressing a happy tone when students get the concept correct.

5.6 Analysis 3: Comparing Sentiment and Tone

In addition to looking at the use of punctuation marks and how it is correlated with student and teachers level factors, in the next step we also want to explore the relationship between the sentiment of these feedback messages with the usage of punctuation marks and other features of student, teacher and problem levels. In this analysis, we drop all the neutral feedback messages and focus on negative and positive sentiment feedback. We perform a separate set of analyses for the sentiment with and without including the math terms, and drop all the neutral feedback messages as we particularly want to focus on negative and positive sentiment and their association with tone. The total number of observations in each of these cases is 65,997 and 75,719 respectively. For this similar to the prior analysis, we first observe a multilevel null model of logistic regression with teacher-, problem- and student-level identifiers at level 2 to understand the higher

level factor that may correlate with the sentiment of feedback messages. In the next step, we perform a multi-level logistic regression model with the sentiment (either positive or negative) as the dependent variable and using different features of student and feedback such as prior knowledge, score on the current problem, assignment completion, and total words used in the feedback, and use of question marks and exclamation marks in the feedback at level 1, and teacher-identifier at level 2.

Table 5.6: The resulting model coefficients for the logistic regression model on the sentiment (positive = 1 and negative = 0) of feedback messages.

	Sentiment w Math		Sentiment w/o Math	
	Variance	Std. Dev.	Variance	Std. Dev.
<i>Random Effects</i>				
Teacher	2.183	1.478	1.976	1.406
	β	Std. Error	β	Std. Error
<i>Fixed Effects</i>				
Intercept	-2.221***	0.099	-1.009***	0.089
Score	2.256***	0.030	1.550***	0.025
Total words	-0.087*	0.035	-0.424***	0.030
Presence of Question marks	0.030	0.049	0.045	0.040
Presence of Exclamation marks	2.465***	0.059	2.317***	0.053
Prior knowledge	1.244***	0.084	0.907***	0.072
Assignment Completion	0.887***	0.051	0.750***	0.047

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$; β denote standardized coefficient

5.6.1 Results of Analysis 3

The null model of logistic regression for predicting sentiment with and without math terms explains 73% and 70% variance of the data respectively. For sentiment with math terms, student-level explains 1.01 variance, problem level explains 3.39 variance, and teacher level explains 4.52 variance. Similarly for sentiment without math terms, the student level has 0.69, the problem level has 2.28 and the teacher level has 3.71 variances. This suggests that some teachers are more likely to use positive sentiment while others use nega-

tive, and for certain problems, students typically see one sentiment versus the others in their feedback. For both sentiments with and without math terms, the results of multi-level logistic regression are presented in Table 5.6. All the fixed effects variables are statistically significant and are correlated to the sentiment except for the presence of question marks. We cannot tell much about the sentiment of feedback solely based on the presence of question marks. If a student gets a high score on a problem they are likely to get positive feedback, and a lower score means more of negative feedback from teachers. Similarly, sentiment is positively correlated with the presence of exclamation marks in the feedback, prior knowledge, and assignment completion. However, the number of words is negatively correlated, and the effect is stronger in sentiment without math terms than with math terms. This could be the effect of dropping some of the math terms when predicting the sentiment in the second case.

5.7 Conclusions and Future Work

This paper aims to explore trends in teacher sentiment and tone when writing feedback messages to students in a mathematics class. We use a generic sentiment analysis method and explore how such methods can be applied to a mathematical context. Through conducted analyses, we find that sentiment and student performance metrics are correlated, but also find potential risks in utilizing pre-trained sentiment models without considering validity within the context of application; in this regard, the use of punctuation actually offers a simpler means of interpreting the valence of teacher feedback when considered in conjunction with provided scores.

The study however has several limitations which should be noted. First, we addressed the issue of generalization of the pre-trained sentiment model by omitting mathematics terms, while future work could focus on retraining or fine-tuning such models for application within mathematics domains. Also in the next steps, we could explore using other ways to measure tone in feedback, through the use of various natural language processing techniques. This work may be further expanded by exploring the use and effectiveness of different feedback writing styles based on tone and sentiment across various students in a mathematics classroom.

AUTO-SCORING STUDENT RESPONSES WITH IMAGES IN MATHEMATICS

Teachers often rely on the use of a range of open-ended problems to assess students' understanding of mathematical concepts. Beyond traditional conceptions of student open-ended work, commonly in the form of textual short-answer or essay responses, the use of figures, tables, number lines, graphs, and pictographs are other examples of open-ended work common in mathematics. While recent developments in areas of natural language processing and machine learning have led to automated methods to score student open-ended work, these methods have largely been limited to textual answers. Several computer-based learning systems allow students to take pictures of hand-written work and include such images within their answers to open-ended questions. With that, however, there are few-to-no existing solutions that support the auto-scoring of student hand-written or drawn answers to questions. In this work, we build upon an existing method for auto-scoring textual student answers and explore the use of OpenAI/CLIP, a deep learning embedding method designed to represent both images and text, as well as Optical Character Recognition (OCR) to improve model performance. We evaluate the performance of our method on a dataset of student open-responses that contains both text- and image-based responses, and find a reduction of model error in the presence of images when controlling for other answer-level features.

Proper citation for this chapter is as follows:

Baral, S., Santhanam, A., Botelho, A.F., Gurung, A., & Heffernan, N.T. (2023). Automated Scoring of Image-based responses to Open-ended mathematics question. In The Proceedings of the 16th International Conference on Educational Data Mining. (EDM 23).

6.1 Introduction

The blending of educational technologies with machine learning and statistical modeling has led to the emergence of tools designed to augment instruction. While some such tools are designed to automate certain tasks for the teacher (e.g. [10, 154, 9]), others attempt to improve the efficiency with which teachers are able to assess student work and write directed feedback to guide learning.

In the context of mathematics education, teachers utilize a range of question formats to assess students' understanding of covered topics. Prior work has described these question types in terms of "close-ended"

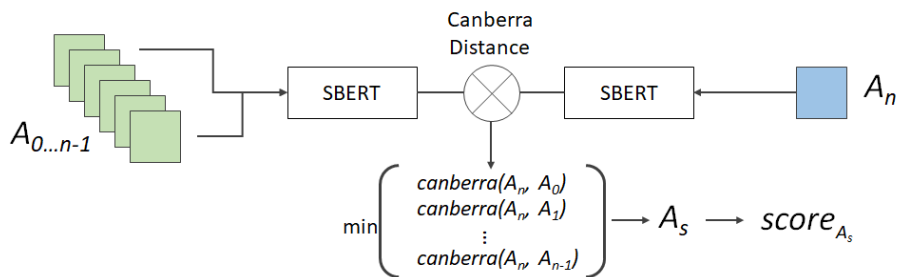


Figure 6.1: Simplified representation of the SBERT-Canberra method to generate a predicted score by identifying the most similar historic response to a given new student answer using Canberra distance within an embedding space.

and “open-ended” problems, distinguishing various types of problems by the difficulty with which answers to such questions may be automatically assessed by a simple matching algorithm. Multiple choice or fill-in-the-blank problems, as examples of close-ended problems, often allow for a small number of acceptable “correct” answers (i.e. in most cases, there is a single answer considered as correct). Although prior works have demonstrated the utility of these types of answers for measuring student knowledge (e.g. the extensive work on knowledge tracing [80, 280]), teachers often rely on the use of open-ended problems to gain deeper insights into the processes and strategies employed by students to solve such problems, as well as their ability to articulate their approach using proper mathematical terminologies. Short answer and essay question types are common in this regard, often with prompts such as “explain your reasoning”, but other open-ended formats are also common in the domain of mathematics.

For mathematics, teachers often rely on the use of visual representations in conveying mathematical concepts. The use of diagrams, number lines, graphs, tables, and sometimes even pictographs are commonly used to portray numerical and algebraic relationships. Just as these are used for instruction, students are also commonly asked to generate these types of visual representations to demonstrate their understanding. While open-ended work has typically referred to the use of text and natural language within prior research (e.g. [113, 395, 30]), the definition extends to drawings and similar artifacts produced by students. Tools such as GeoGebra[165] and Desmos[110] are examples of computer-based applications that allow students to interact with graphs and algebraic expressions. While tools like these exist, many teachers still prefer to use more traditional technologies, often in the form of paper and pencil or other physical media (e.g. blocks) in conjunction with computer-based technologies; some systems encourage this blending of media by allowing students to take pictures of their work and upload them as responses to open-ended problems.

This paper builds on prior work which focused on the development of an automated scoring tool for student answers to open response problems in mathematics [30]. Baral et. al, reported on how many student

responded to open-response problems with images of their work (in the form of written mathematical equations and expressions as well as drawings of graphs, number lines, and other visual representations), whereas several others preferred to respond with a combination of an image of their work combined with a typed textual explanation within a single student response (e.g. the student draws a graph, uploads the image and then types a description of their thought process with the image of the graph). These cases were, unsurprisingly, found to contribute significantly to the model error as the presence of images in student responses were not previously accounted for within the developed methods. This work seeks to take initial steps toward understanding how recent advancements in areas of deep learning-based image and text embedding methods may help to address these challenges.

Specifically, this paper addresses the following research questions:

1. Does the use of pre-trained deep learning image and text embedding methods lead to improved performance in the context of previously-developed open response scoring models?
2. Are there differences in terms of the resulting model performance when comparing across different types of image-supporting embedding methods?
3. Does the incorporation of image-supporting embedding methods reduce the correlation between the presence of images in student responses and modeling error when accounting for other answer-level covariates?

6.2 Related Works

6.2.1 Automated Scoring Models

With the development of online learning platforms, there has been a growing body of research in the development of automated methods of assessment for analyzing and providing immediate feedback on students' work. These developments have prevailed in multiple domains of science [205, 39], programming [235, 286, 382], writing [188, 7, 57, 306, 397], mathematics [199, 113, 30] and college level courses [95]. In the domain of mathematics, auto-scoring have been developed for closed-ended problems with single or limited correct answers (e.g., multiple-choice question, fill-in-the-blank, check all that apply) [10, 154] to more open-ended problems with multiple possible solutions (eg. short answer, long answer, Explain in plain english.) [199, 113, 126, 30, 395, 396, 31, 332]. Some of these works support pure mathematical content [199], while others support combination of both mathematical and textual answers [113, 30, 31, 396]. However, most of these auto-scoring methods in mathematical domains are limited to either text or mathematical content, and

a very few have started focusing on automating responses for image-based responses.

6.2.2 *Methods for Image Analysis and Representation*

Optical Character Recognition (OCR) is an extensive field of research in image processing, that explores the recognition and conversion of handwritten textual information to machine-encoded text, such that this information could be further processed and analyzed. Studies such as Shaikh et al. (2019) [331], utilizes OCR-based methods, combined with Convolutional Neural Networks(CNN) in auto-scoring structured handwritten answer sheets of multiple choice questions. Other studies like [348] propose an automated scoring system for handwritten student essays in reading comprehension tests, utilizing handwriting recognition and machine learning-based automated essay scoring methods. Khuong et. al [185] in their work proposes clustering handwritten mathematical answers scanned from paper-based exams, to improve the efficiency of human raters in scoring these answer sheets. Another study from Gold et. al [131], in their attempt to auto-score handwritten answers, presents the challenges of using handwriting in intelligent tutoring systems. Further, they present, how the lack of better recognition systems in these cases leads to poor scoring performances.

Recent advancements in the areas of deep learning and computer vision have led to the development of large-scale models of image representation and classification. ImageNet [88] is a large-scale image dataset widely used for training and evaluating computer vision models. Trained over 14 million images belonging to more than 22,000 different classes, ImageNet is considered a benchmark for image classification tasks. CLIP (Contrastive Language-Image Pre-training) [298] is a recently introduced image classification model based on transformer architecture, commonly used in natural language processing tasks. This method is able to encode both natural languages (text) and images in the same vector space by using a multi-modal pre-training approach. The proposed methods in this work utilizes the CLIP model to represent image and text-based answers.

6.2.3 *The SBERT-Canberra Model*

This work utilizes an auto-scoring method developed through several prior works [30, 47], referred to as the SBERT-Canberra model. As illustrated in Figure 6.1, the method produces a predicted score, $score_{A_s}$, for a new student answer, A_n , by leveraging the single-most-similar historic student answer, A_s . The method utilizes Sentence-BERT [302] to first generate a 768-valued feature vector for both A_n as well as all teacher-scored historic student answers, $A_{0..n-1}$ before then making a full pairwise comparison of A_n to these historic answers using Canberra distance[178]; Canberra distance is a rank-order-based distance measure

that was found to more closely align to how teachers identify similarity in comparison to other distance measures such as Euclidean and Cosine Similarity [47]. From this, A_s is identified and its teacher-given score is used as the prediction for A_n ; the method, therefore, adopts a variation of K-Nearest-Neighbors and has exhibited notable performance when evaluated compared to a range of baseline models [30, 113], despite its simplicity.

Through prior work, several weaknesses of the auto-scoring method have also been identified by means of a multi-level regression-based error analysis [30]. From this, four primary areas of weakness were identified: 1) model error varied greatly from problem to problem, 2) there seemed to be variation in teacher grading, 3) the presence of numbers, expressions, and equations in textual explanations correlated with higher error, and 4) the presence of images in student answers correlated with higher error. Subsequent follow-up works have explored three out of these four weaknesses, examining how answers from similar problems can be leveraged to improve predictive power for problems with smaller sample sizes [308], explore the contextual factors that contribute to variance in teacher grading practices [144], and leverage the most-frequent mathematic terms, numbers, and expressions to reduce modeling error [31]. Following these works, this paper seeks to address the fourth weakness by exploring potential methods of representing both textual and image data within similar embedding spaces.

6.3 Dataset

In this study, we utilize a dataset of student open-ended answers in mathematics from the prior studies [30], to compare directly with the prior works. This dataset consists of 150,477 students' answers to 2,076 different open-ended mathematics problems and scores given by 970 different teachers to these responses. The scores given by teachers to these responses are on an ordinal 5-point scale ranging from 0 to 4. The student responses given to these math-based questions are typically seen as a combination of textual responses (typed directly into the learning platform), mathematical expressions and equations, and images uploaded as a part of their work. The current dataset includes 3712 image responses in total to 311 different math problems. Some example image responses given by students are presented in Figure 6.2. As seen from these examples, the image-based student answers are of different types – some are handwritten, whereas others are digitally drawn images. In addition to this, these images can include handwritten text, diagrams, and graphs on a piece of paper. We can see lots of variations in these responses, in both text and image format.

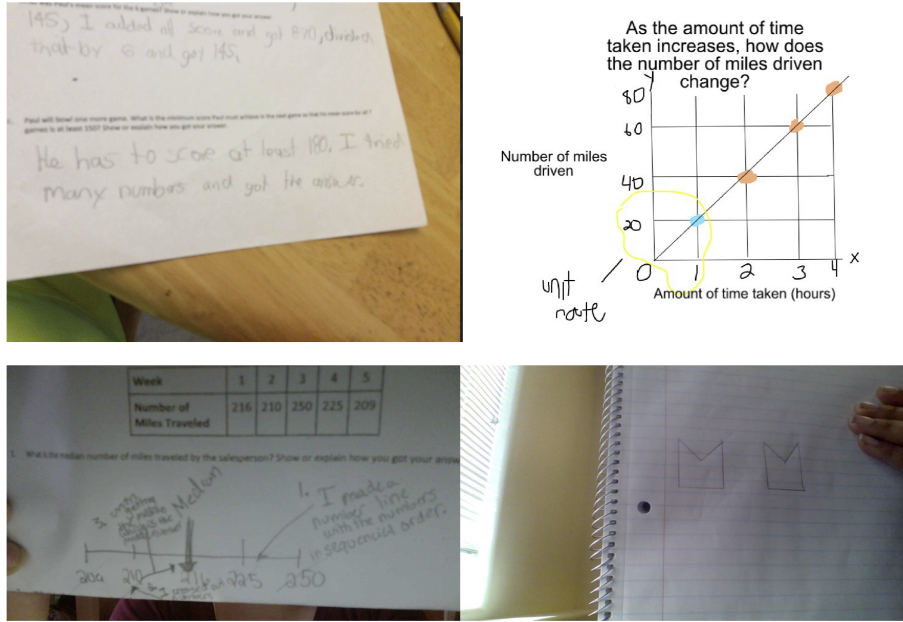


Figure 6.2: Examples of image-based responses from students given in response to Open-ended math problems

6.4 Methodology

Utilizing the dataset from [30] and a similar model design to auto-scoring student open-response answers, we propose an extension to this prior work to support image-based responses. Similar to [30], we train a separate model per problem and perform a 10-fold cross-validation for training. For the problems without any training data, a default model based on word counts, trained across all problem data is used similarly to the prior works. In this paper, we explore and compare three different methods which we describe in detail in the following sections.

6.4.1 CLIP-Text Method

As stated earlier, the prior works [30], is a similarity ranking-based method, that first converts each student's answers to a 768-valued vector representation using Sentence-BERT[302], and compares answers using this vector representation and Canberra distance[178]. In our current method, we use a similar model structure with a different embedding method. This method is based on CLIP (Contrastive Language-Image Pre-training)[298] for encoding textual responses.

In the first method which we call the 'CLIP-Text' Method, we perform a text comparison similar to the prior SBERT-Canberra model, without accounting for image-based responses. Using the CLIP[298] model, we first embed the textual responses ignoring all the image responses. For any new answer in the test dataset,

we compare them with the training set, by first generating a vector representation, and then comparing the vectors using Canberra distance to find the most similar pair of text responses. Using the most similar text, we utilize the score given by teachers to this similar response, in suggesting a score for the new response. In the CLIP-Text Method, we ignore the images, as we want to see how well the CLIP model does with just the text responses to directly compare it to the prior method. For any empty student responses, the model assigns a score of '0', and also for responses with no textual answers (images are discarded in this method, so if a response contains only an image, it is assigned a score of 0).

6.4.2 CLIP-Image Method

The second method which we call 'CLIP-Image' method, addresses both images and text in student responses. This method is similar to the 'CLIP-Text' method, with the addition of image embeddings in comparing the similarity of responses. The CLIP model uses separate text and image encoders and allows embedding text and images into the same vector space. With the CLIP model, we first encode textual and image responses into a vector representation. If a student response contains both text and images, the text part is discarded and just the images are encoded in this method. Once all the responses in the training data are encoded, for a new student answer (with either image or text-based response), its corresponding encoding is calculated and compared to the embeddings in the training data, and the most similar response is selected based on the shortest Canberra distance between the new response and the responses in the training set.

6.4.3 CLIP-OCR Method

The third method is called 'CLIP-OCR' method which is based on state-of-the-art Optical Character Recognition (OCR). This method uses the Tesseract engine[346] from Google for text extraction. Tesseract is an open-source OCR engine, that extracts both printed and written text from images. Similar to the 'CLIP-Text', this method, then encodes the original textual responses, and also the extracted text from images (without completely ignoring the image responses). The text information from the responses is then encoded using the CLIP model, and finally, any new response is compared to the historic responses in the training data using the encodings and Canberra distance, to get a score prediction.

6.5 Results

To compare the current approaches directly to the prior methods from [30], we utilize similar evaluation methods, using a Rasch model. The use of the Rasch model allows a fairer comparison that accounts for

Table 6.1: Model Performance compared to the auto-scoring methods developed in the prior works [30]

Model	AUC	RMSE	Kappa
Current Paper			
Rasch* + CLIP-Text	0.852	0.594	0.469
Rasch* + CLIP-Image	0.854	0.587	0.471
Rasch* + CLIP-OCR	0.854	0.588	0.471
Prior works[30]			
Baseline Rasch	0.827	0.709	0.370
Rasch* + Random Forest	0.850	0.615	0.430
Rasch* + SBERT-Canberra	0.856	0.577	0.476

*These rasch models also included the number of words.

factors external to the observed student response, such as student ability and problem difficulty. We evaluate the methods using three different metrics – AUC score, Root Mean Squared Error (RMSE), and multi-class Cohen’s Kappa. The AUC score here is calculated as an average AUC over each score category and Root Mean Squared Error(RMSE) is calculated using the model estimates as a continuous-valued integer scale. The results of three methods as compared to the prior works [30] are presented in Table 6.1.

The result suggests that the CLIP-Text that uses the sentence embeddings from OpenAI CLIP model [298] has an AUC score of 0.852, RMSE error of 0.594, and Kappa of 0.469. Though the model doesn’t outperform the prior SBERT-Canberra method [30] of auto-scoring, the difference in each of the scores is very small. The next method CLIP-Image, which compares both sentence and image embeddings using the OpenAI CLIP model, outperforms the CLIP-Text method across all three evaluation metrics used (though the difference in these scores is minimal). This method has an AUC score of 0.854, RMSE error of 0.587, and Kappa of 0.469. The next method CLIP-OCR, based on text extraction from images using OCR methods, has a similar performance to the CLIP-Image model. Though the newly introduced methods do not outperform the prior text-based method, the introduction of auto-scoring image responses is something novel that this work explores. And we can see improved performance with the addressing content from image-response in the CLIP-Image and CLIP-OCR model, than solely using text-based responses in the CLIP-Text model.

Table 6.2: The resulting model coefficients for the linear regression model of error for the auto-scoring method, conducted as a part of the error analysis similar to the prior method from Baral et. al [30].

	CLIP-Text		CLIP-Image		CLIP-OCR	
	B	Std. Error	B	Std. Error	B	Std. Error
Intercept	0.356***	0.006	0.324***	0.006	0.324***	0.006
Length of Answer	0.002***	0.000	0.002***	0.000	0.002***	0.000
Avg. Word Length	0.014***	0.001	0.016***	0.001	0.016***	0.001
Numbers Count	0.0001***	0.000	0.0001***	0.000	0.0001***	0.000
Operators Count	-0.001**	0.000	-0.001***	0.000	-0.001***	0.000
Equation Percent	0.161***	0.009	0.211***	0.009	0.208***	0.009
Presence of Images	2.432***	0.019	0.496***	0.018	0.585***	0.018

*p < 0.05 **p < 0.01 ***p < 0.001;

6.6 Error Analysis

As previously introduced, prior work conducted an error analysis to understand the limitations of the SBERT-Canberra method [30]. This error analysis involved the calculation of several student answer-level features and using a linear regression analysis with the absolute prediction error (absolute difference between the teacher-provided score and the prediction from the model) as the dependent variable. This analysis reported that the largest amount of error in the SBERT-Canberra model was correlated with the presence of mathematical terms and equations and the presence of images in the answer text.

In this paper, we propose a method to auto-score responses in presence of both text and images. Although the proposed methods do not outperform the previous method on auto-scoring strictly text-based answers, we hypothesize that this could be a result of using a different method of embedding text; there may be an inherent trade-off where performance is reduced for textual responses but results in improved performance where there are images (averaging out to little-to-no overall improvement). Also, from the results, we have seen improvements in the performance of the ‘CLIP-Image’ and ‘CLIP-OCR’ methods (that addresses the content of the image when auto-scoring) over the ‘CLIP-Text’ method (which is just based on text responses). To further study the factors that contribute to the error of these models, and to verify whether introducing image components in the text-based models actually improve the performance in presence of images, we replicate the error analysis from Baral et. al [30]. Using features from student answers including ‘Length of answer’,

'Average word length', 'Total numbers count', 'Total operators', 'Percentage of equations' and 'Presence of images' as the dependent variables and Absolute model error as the independent variable, we perform three different linear regression analyses corresponding to the three proposed methods for auto-scoring.

6.6.1 Results of Error Analysis

The results of the error analysis are presented in Table 6.2. All the features from student answers are statistically significant in predicting the modeling error in all three proposed methods. However, most of these features have low coefficient values, suggesting a relatively small effect, with the exception of 'Equation Percent' and 'Presence of Images' which are positively correlated with the model error in all three cases. This is similar to the results of error analysis from prior study [30]. For the 'CLIP-Text' model, the coefficient for the presence of images is 2.432, suggesting that the presence of images in answers attributes to a notable amount of error in the model prediction, even when considering the difference in feature scaling. However, the coefficient value decreases to 0.496 in the 'CLIP-Image' method, and 0.585 in the 'CLIP-OCR' method. This decrease suggests that the introducing image component to the 'CLIP-Text' method using embedding and OCR-based text extraction actually helped the model improve in presence of images. It is also important to note that this work does not explicitly address mathematical terms (including numbers, expressions and equations) in the score prediction as has been suggested by other work [31]. Also, we see an increase in the coefficient values for equation percentage from 'CLIP-Text' to 'CLIP-Image' and 'CLIP-OCR'. For the 'CLIP-Text' method, we discard any images from the answer text, whereas for the other two methods, if there is a response that contains both image and text we discard the text from these responses and just consider the images. The change in the coefficient values for equation percent could be a result of this quality.

6.7 Limitations and Future Works

This paper represents an initial step toward improving state-of-the-art methods for auto-scoring student responses to mathematical problems in presence of images. This is a preliminary work conducted towards exploring the feasibility and challenges in auto-scoring student image responses in the mathematical domain. Thus, the methods presented have several limitations and challenges that can be addressed with future work.

The proposed methods in this work use CLIP model [298] trained on a large variety of datasets of images and natural language available over the internet. While this method shows promising results in recognizing a range of common objects, the pre-trained model may not have been exposed to the dataset of student handwritten or hand-drawn mathematics; the model was trained for application in very broad domains to recognize

objects and is not optimized for identifying similar responses on paper. It has also been found that while the CLIP model learns a capable OCR system, it exhibits low accuracy in the case of handwritten digits in the widely-used MNIST dataset [298]. Further, fine-tuning this model on a mathematical dataset could lead to better model performance.

It is also important to note that the OCR method is based on the Tesseract [346] engine; this is known to be sensitive to poor quality images, complex backgrounds, variation in the handwriting styles and ambiguity in the characters [346]. All of these are the common qualities of the images found in our dataset. While this method supports digital images (that are screenshots of work done on a computer), the method has low accuracy in extracting textual information from handwritten answers. Thus, exploring better OCR methods that support both handwritten and digital textual answers would better improve these auto-scoring methods for images. Further, both of the proposed methods that support images, inherently discard the additional text if present in the response. These texts may present additional supporting information to the image-based answers, so it is important to explore how to address this when evaluating these responses.

Apart from the limitation mentioned above, the process of analyzing and processing these image-based answers in itself is a challenging task, as we can see a lot of variation in these images of student-provided answers. Figure 6.2, presents some examples of image-based student answers. The student work in these images are not always clearly presented and structured – some handwriting is hard to read, the images sometimes are of low resolution and are blurry, the use of pencils makes the writing feint and hard to read, and lacks consistent formatting. Due to the freedom provided to students by the use of paper and pencil to draw out their solution, the resulting answer is not always structured in the same way from student to student. Future work could help address some of these challenges by implementing a more rigorous cleaning and preprocessing procedure prior to applying any image representation models. Cropping images to focus on the prominent aspects of student work, rotating images to improve the consistency of orientation, and even color correction can help improve the clarity of the work.

In all of this work, there are also several ethical concerns that should be considered in developing and applying these various methods. Images may contain Personally Identifiable Information (PII) such as students' names, faces, skin color, etc. which exposes a potential risk of biases or disparate performance in regard to the machine learning models. Future works could mitigate some of these challenges by utilizing some of the pre-processing methods described above, but also emphasizes the importance of evaluating these scoring models for potential biases or unfairness in their predictions.

6.8 Conclusion

In this study, we have presented preliminary work towards developing an auto-scoring method for student response in mathematics that includes images. By building upon the prior research in auto-scoring text-based mathematical answers, we have proposed methods for representing and scoring image-based responses. While our proposed methods did not outperform the current state-of-the-art approach for auto-scoring, they showed comparable accuracy across all three evaluation metrics used. The results of the conducted error analysis further indicate that using pre-existing methods of text and image embeddings can enhance the performance of the auto-scoring models in presence of images.

Our findings from this study points toward new directions for research in the area of analyzing and processing image-based student responses in mathematics.

Chapter 7

CONSIDERATE, UNFAIR, OR JUST FATIGUED? EXAMINING FACTORS THAT IMPACT TEACHERS

It is particularly important to identify and address issues of fairness and equity in educational contexts as academic performance can have large impacts on the types of opportunities that are made available to students. While it is always the hope that educators approach student assessment with these issues in mind, there are a number of factors that likely impact how a teacher approaches the scoring of student work. Particularly in cases where the assessment of student work requires subjective judgment, as in the case of open-ended answers and essays, contextual information such as how the student has performed in the past, general perceptions of the student, and even other external factors such as fatigue may all influence how a teacher approaches assessment. While such factors exist, however, it is not always clear how these may introduce bias, nor is it clear whether such bias poses measurable risks to fairness and equity. In this paper, we examine these factors in the context of the assessment of student answers to open response questions from middle school mathematics learners. We observe how several factors such as context and fatigue correlate with teacher-assigned grades and discuss how learning systems may support fair assessment.

Proper citation for this chapter is as follows:

Gurung, A., Botelho, A.F., Thompson, R., Sales, A.C., Baral, S., & Heffernan, N.T. (2022). Considerate, Unfair, or Just Fatigued? Examining Factors that Impact Teachers. In Proceedings of the 30th International Conference on Computers in Education (ICCE 2022).

7.1 Introduction

In the context of education, a significant amount of research has been devoted to identifying, examining, and mitigating risks that particular policies, interventions, and instructional strategies may introduce as to the types of opportunities made available to students. Particularly with the introduction of computer-based learning platforms (CBLP), researchers are able to explore issues of fairness and bias through data-driven methods.

Traditional assessments are conducted in two formats: subjectively and objectively. The rise in the integration of technology in classrooms has facilitated the growth of CBLPs. Various CBLPs have been de-

veloped with the goal of alleviating difficult or tedious tasks faced by teachers. The most notable of the functionalities is the automation of objective assessments. Common in numerous contexts, the use of close-ended questions such as multiple choice and fill-in problems can be easily automated by computers where there are traditionally a small finite number of acceptable correct answers. The use of such questions has allowed developers to expand upon assessment processes to enrich the learning experience by offering additional feedback and on-demand help [5, 4, 261, 274]. However, implementing subjective assessments has been more complicated due to its dynamic nature. It is challenging to extend the same type of support to open-ended problems such as short answers and essay problems. While recent advancements in natural language processing (NLP) and machine learning have made progress towards automating the domain of subjective assessment, the task of assessing open-ended student work still remains predominantly a manual task for teachers.

Writing is a critically important skill that helps teachers understand their students' thought processes and the ability to formulate arguments and justifications for their work [40, 135, 369]. In the domain of mathematics, on which the analyses of this work focus, teachers commonly use open-ended problems to gauge student knowledge as close-ended problems can often be solved by shallow learning and applying procedural rules [338, 215]. While subjective assessments are highly valuable, automating the process is not without risk; they are more dynamic compared to objective assessments. The dynamism is due to the variance in responses and the incongruity in teacher grades; the incongruity is caused by various intrinsic and extrinsic factors. Prior work has found biases in grading behavior attributed to the "Halo Effect" [76], characterized by a judgment made based on an attribute or characteristic of an individual; commonly, such attributes include gender [347, 309], ethnicity [379, 115], name or surname [203], and report of gifted status [19]. Furthermore, researchers exploring the Halo Effect found the initial favorable impression of students influenced their later evaluation [220, 221]. While the use of rubrics or other standardized procedures helps to evaluate students along with common sets of metrics, the ultimate grade is typically based on how well a teacher has judged the student to demonstrate sufficient knowledge of the given topic. It is important to emphasize that measured biases do not necessarily equate to unfair assessment or evaluation as a teacher's knowledge of their students can also be very positive in terms of providing individual support through feedback and other communication [161, 159, 176]. In approaching assessment, teachers may consider a number of contextual factors when evaluating student work. In light of this, several questions emerge in terms of how assessment should be conducted to ensure fairness among students, particularly as researchers are moving to develop automated methods that attempt to mimic teacher grading practices.

Our goal in this work is to explore teacher assessment pertaining to open-ended questions. Using data collected from students working in a learning system in pre-COVID-19 classroom settings, we report on a pilot study to examine and explore teachers' approach to assessing open-ended student work and how knowledge of students may be considered in the grading process. We build on the study by conducting an exploratory analysis examining whether student-level attributes are predictive of teacher-provided grades when controlling for answer-level descriptors. Finally, we explore whether teacher grading fatigue poses risks to the fairness of student assessment.

In consideration of exploring factors that may affect fair student assessment, this paper addresses the following research questions:

- Do teachers grade students differently when the students are anonymized?
- When controlling for answer-level features, are factors of prior student performance and effort a reliable predictor of teacher-provided assessment scores?
- Does the order in which teachers assess students appear to impact their grading?

To address these questions, we conduct 3 studies exploring various factors that are likely to affect or potentially bias teacher assessment. In the first study, we examine whether the anonymization of student identifiers affects how teachers score their own students. In the second study, we conduct a regression analysis to examine how measures of knowledge and effort correlate with teacher-provided assessment scores for student open-ended work. Finally, we explore the potential effects of grading fatigue, expressed through the ordering in which student responses are graded, on teacher-provided scores.

7.2 Background

Growth and innovation in Educational Technology (Ed-Tech) have broadly influenced the adaptation and regular usage of CBLPs in classrooms. Researchers and developers approached the design of learning platforms to consider students' various learning needs [135, 193, 299] to developing generic platforms [16, 57, 79, 154]. Systems developed for such content areas as writing skill [7, 57, 61, 314], mathematics [16, 154], and programming [286, 382] are among the many examples of developed learning systems. Within these systems, a variety of features and supports are commonly developed to support different aspects of learning, including the crowdsourcing of problems and solutions [39, 94], and the availability of hints and explanations [384, 63, 184].

Perhaps the most prominent feature of CBLPs is the ability to offer immediate feedback to students. Traditionally assessments are administered in objective and subjective forms. In many domains such as mathematics, students typically work through close-ended problems that can be assessed by simply comparing student answers with a finite set of acceptable correct responses (often with a simple “exact-match” approach), but across domains, the use of open-ended questions, allowing students to utilize language to explain their reasoning, have made it more difficult for CBLPs to immediately score. While examples of automated scoring tools exist [30, 113, 7, 57, 188], many CBLPs still rely on manual assessment of open-ended student responses by teachers. A primary challenge with subjective assessment associated with this manual grading process is its susceptibility to bias.


As introduced in the previous section, the Halo Effect has been the focus of prior research within the context of subjective assessment of student work [253, 252]. Prior research have found stereotyped biases interacting with gender [347, 379, 309, 224], ethnicity [279, 379, 115], “likeability” and attractiveness [200, 64], student names [203], and perceived ability [20, 19]. Other studies exploring the Halo Effect found effects persisting across multiple assignments from the same student [89], and that this effect was specifically identified in cases where teachers were assessing student writing samples [123]. One possible solution proposed to mitigate teacher bias is the anonymization of student identity during subjective grading [220, 221], motivating the current work.

7.2.1 Open-Ended Problems in ASSISTments

In this paper, we analyze the teachers grading open-ended mathematics problems in ASSISTments. ASSISTments is a CBLP that allows teachers to assign content (primarily in middle-school mathematics) and monitor student progress. The system provides students with immediate correctness feedback on close-ended problems and offers computer-provided help in the form of on-demand help and scaffolding. Open-ended problems in ASSISTments are available in two different structures: (a) primary problem or (b) sub-problem of a multipart problem. Figure 7.1(a) shows an open-ended problem presented as the main problem, and figure 7.1(b) shows an open-ended problem presented as a subpart of a multipart problem. A multipart problem can have more than one open-ended sub-problem as well.

Once students have completed their assignments, teachers can grade each student’s response for each problem. Figure 7.2 shows the interface teachers can use to grade their students’ responses and provide feedback. Teachers have the option to anonymize their students during the grading process where their identity is hidden, and the student responses are shuffled, but the page defaults to showing all of their students’

What is something that is *definitely* true about the value of x ?

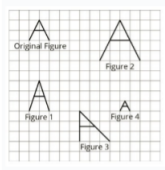


Type your answer below:

© 2019 Illustrative Mathematics (CC by 4.0) Full Attribution Copied for Free From Open Up Resources (CC by 4.0) (a)

Part A

Here is a figure that looks like the letter A, along with several other figures. Which figures are scaled copies of the original A?



Select all that apply:

Figure 1

Figure 2

Figure 3

Figure 4

Part B

Explain how you know.

Type your answer below:

© 2019 Illustrative Mathematics (CC by 4.0) Full Attribution Copied for Free From Open Up Resources (CC by 4.0) (b)

Figure 7.1: The different types of open-ended problems on ASSISTments. (a) the main problem is open-ended (b) a multi-part problem where the second problem is open-ended.

identifying information.

Student	Response	Score	Teacher Comment
Student Name 1	The square root of 9 is 3	▼	<input style="width: 100%; height: 20px;" type="text"/>
Student Name 2	written response	▼	<input style="width: 100%; height: 20px;" type="text"/>
Student Name 3	The area of Q is 45. You have to divide it by the area of P which is 5. The answer is 9. Then, we need to take the square root of 9 and that is our scale factor.	▼	<input style="width: 100%; height: 20px;" type="text"/>
Student Name 4	the area of polygon q is 45 aquare units. the area of polygon p is 5 square units. that means the scale factor is 9 square units. the square root of 9 is 3 scale units.	2 ▼	<input style="width: 100%; height: 20px;" type="text"/>
Student Name 5	find the area of q which is 45 then do 45 divided by 5 which is 9 then you find the square root of 9 which is 3	4 ▼	Great Work!
Student Name 6	I know this because the area of Q is 45, and when you compare the shapes, you an see that Q is nine times bigger. So if you square 9, you get 3 and thats your scale factor.	3 ▼	squaring 9 is not 3

Figure 7.2: The interface for teachers to grade students' responses on open-ended problems.

7.3 Study 1: Examining Grading Differences When the Student is Anonymized

Our first analysis explores whether teachers assess their students differently when they know the student's identity compared to when students are anonymous. In a purely unbiased, impartial scenario, teacher grading

behavior would be solely dictated by the quality of the response as determined by how well the student was able to articulate their thoughts and demonstrate their knowledge of the concepts; this may include aspects such as grammar, use of mathematical terms (in mathematics contexts), and overall completeness. However, we posit there may be other factors that teachers consider including effort and prior performance. Even from a motivational perspective, a teacher may be inclined to bias their grade in a positive direction for students who normally under-perform but applied notable effort as a way to encourage similar behavior in the future (i.e. an arguably positive example of bias). The danger, however, is in whether these perceptions, implicit or explicitly applied during the grading process, impact the types of opportunities that may be available to a student.

We conducted a study involving 9 teachers who commonly assigned and graded student open responses in ASSISTments (14 teachers were initially involved as part of a larger study, but only 9 participated in this portion). We selected 3 problems containing an open-response sub-part that was assigned by teachers within the month prior to beginning the study. Each teacher had already assigned and graded their student work for at least one of the three problems using the default scoring paradigm (i.e. teachers scored while knowing the identity of each student). In the month after this scoring was completed, we collected all student responses from across all teachers, anonymized them, and then randomly selected approximately 25 student responses to present to each teacher, ensuring that at least 10 of the responses were from each respective teacher's own students (if fewer than 10 were randomly selected, the difference was sampled from that teacher's student responses and added to the 25 given to the teacher); duplicate responses (e.g. empty responses or answers of "I don't know") were removed so that each teacher had a set of unique answers (resulting in some teachers having fewer than 10 responses from their own students if one was a duplicate). Each teacher was then asked to grade the given set of student answers, such that, for their own students, they would be anonymously grading the same set of answers as they had non-anonymously in the past (the additional responses and amount of time between grading reduced the likelihood that teachers would recognize their own students' responses). This presented the opportunity to measure each teacher's intra-rater reliability (i.e. how well they agreed with their past selves) and whether their grading was biased in a particular direction when they knew the student compared to not.

As the grading was done on a 5 points scale of 0-4; we applied Weighted Cohen's Kappa with linear weights to measure the variation in teachers grading behavior per response and found that the agreement coefficient as low as $k=0.2293$ and as high as $k=0.7368$, indicating that there was a large degree of disagreement between the two-time points, as shown in Table 7.1. These resulting scores were notably lower than we had

Table 7.1: Exploring the grading behavior of teachers when they had access to students' identity vs. when students were anonymized using Linear Weighted Cohen's Kappa.

Teacher	N	Intra-rater reliability (Weighted Cohen's Kappa)	Intra-rater reliability (Relaxed Cohen's Kappa)	Avg. grade diff (initial - anonymized)
Teacher1	10	0.3878	0.8077	-0.2
Teacher2	10	0.7368	1	0.2
Teacher3	10	0.3939	0.6269	-0.2
Teacher4	10	0.4737	0.7368	0.3
Teacher5	11	0.2293	0.443	0.27
Teacher6	19	0.3596	0.5662	0.57
Teacher7	9	0.3571	0.3571	0.44
Teacher8	10	0.4531	0.5312	0.3
Teacher9	9	0.6539	0.761	-0.66

initially hypothesized, suggesting that there were large differences in how teachers approached the grading of these when students were anonymized.

In addition to the Kappa measure, we also observed a relaxed calculation of Weighted Kappa. Given that grades are given on a 5-point scale and a teacher's assessment may reasonably vary by a small degree, we observe intra-rater agreement with an off-by-one adjustment (e.g. a scoring difference of one in either direction is treated as the same score when calculating Kappa). This adjustment resulted in notably higher Kappas, suggesting that the overall difference of scores was not as large as the first Kappa value suggested; while there was low precise agreement, teachers were relatively consistent within a grade point of themselves when the student was anonymized.

Pairing these Kappa scores with the average grade difference (the right-most column of Table 7.1), we see that there is apparent bias in a particular direction. The positive difference exhibited by the majority of teachers suggests that teachers were more likely to grade anonymous students lower, on average, than when they knew the students. We conducted a permutation test across teachers to estimate the average difference in teacher grading behavior when students were anonymous versus not. We observed that teachers on average were more likely to give higher scores, 0.163 with 95% CI=[-0.1367, 0.4627] when students were not anonymized; while not statistically significant, likely due to our small sample size, this result is suggestive that there was some bias observed in the study.

We followed this empirical analysis with a set of semi-structured qualitative interviews first with the teachers as a group, and then with 2 teachers individually for an extended session to gain better context as to

how they approach grading and why they believed their grades changed during the study. Overall, the teachers unanimously described considering contextual information of the student when approaching the grading process. Several teachers mentioned that they consider motivational aspects (e.g. trying to encourage students to apply effort) when determining grades, as we had initially hypothesized. Several teachers also mentioned attempting to consider student effort when grading but did not use action-level reports in the system to do so (suggesting that it would be too time-consuming). The interviewed teachers similarly acknowledged potential risks to fairness highlighted by the observed differences, with one teacher expressing the intention to always grade anonymously following the study; others disagreed with this course of action citing perceived benefits of understanding the context of student work in order to provide better feedback to students in conjunction with the grade. While arguably anecdotal due to the limited sample size of teachers involved in this study, these conversations highlight the presence of bias in the form of a Halo Effect where teachers have an inherent motivation to give higher grades to perceived lower-performing students (again, not necessarily in a manner that affects fairness, but is still a form of bias).

7.4 Study 2: Exploring Related Factors of Student Assessment

Building on our findings from the pilot study we conduct a quantitative analysis investigating the role of student identity in the grading behavior of teachers. In this section, we attempt to explore the relationship between various answer- and student-level features and teacher-provided grades. Specifically, we seek to address the second and third research questions by exploring 1) whether prior student performance is a strong predictor of grade after accounting for concept-knowledge and other answer-level features (such as, for example, the number of words in the response), and 2) whether a measure of student effort correlates with the grades they ultimately receive while controlling for other measures of knowledge and ability. These analyses are meant to collectively provide insights into what a teacher may consider when assessing student work.

7.4.1 *Description of the Dataset*

We collected a dataset of authentic student responses to open-ended problems from the ASSISTments platform. The dataset consists of action logs of students interacting with open-ended problems for the academic years of 2018 through the beginning of 2020 (i.e., up to but excluding the period of remote learning in response to the COVID-19 pandemic). While not explicitly used as a filtering criterion, a significant portion of the open-ended problems in the dataset are from the OER curricula of EngageNY, Illustrative Mathematics, and Utah Math.

The dataset contains action logs for open-ended problems assigned in the system. Overall, the dataset includes 344,847 action logs from 7,535 students working on 2,268 problems within 2,636 assignments. It is important to highlight that it is additionally the case that problems can contain multiple parts; particularly in the OER content, it is common for open-ended questions to exist as a sub-part of a multi-part problem (e.g. asking students to explain their reasoning after solving a closed-ended question of the same concept). Students worked on 3,404 distinct open-ended problems, reflecting that many problems contained open-response questions for multiple sub-parts.

As in the previous analysis, grades for the student responses observed in this work followed a 5-point integer scale ranging from 0-4. While ASSISTments allow teachers to alter this default grading scale, few teachers change this setting or deviate from using integer values. For our analyses, any grades that do deviate from this are normalized (between 0-4) and rounded to the nearest integer value. Following other work that observed student response time as a measure of effort (c.f. [142]), we use student action logs to calculate how much time they spend while formulating their open response answers as a measure of effort. ASSISTments record three types of actions that are of interest to us for our analyses: starting a problem, leaving without answering to resume later, and submitting a response. We combine these actions into action pairs to compute the amount of time a student spent formulating their response to the problems, accounting for cases of students leaving and resuming work on the problem. These action pairs can be described using the notation of “(first action, second action)” where they represent two consecutive actions of a student taken within a session. The time for the action pairs represents the amount of time a student took between the two recorded actions. Our dataset observed two primary action pairs: “Problem Started-Submitted Response” and “Problem Resumed-Submitted Response” distinguished by the action observed prior to students submitting their response.

With this measure of student time-on-task, we apply a log transformation to create a pseudo-normal distribution and remove samples with a z-score value outside the range [-3, 3]; this filtering step attempts to remove very large outliers that may impact or bias our results. We also examined the open response problems and found that teachers graded only 19,446 (20%) of the 97,105 problems. The resulting final number of action pairs used in our analyses for graded open responses are in table 7.2.

There are several features of student answers that are likely to correlate with teacher-provided grades as identified in Baral et al. (2021) [30]. These features include the “Response Category,” a categorical variable that indicates that the student response contains only words (positive class) as opposed to a mixture of linguistic and mathematical terms and expressions (negative class), as well as the “Number of Words,” a simple count of the number of words (as denoted by spaces) in the student’s answer.

Table 7.2: Filtered Action Pairs of Graded and Ungraded Student Responses. This table shows the dataset after filtering for instances where a student either initiated and completed a problem, or resumed a previously incomplete problem and made a submission.

Action pairs	Graded Responses	Ungraded Responses
(Problem Started, Submitted a response)	18295	73176
(Problem Resumed, Submitted a response)	1151	4483

In addition to the answer-level features, we calculate several student-level features to describe recent and historic performance measures. These measures include “Prior Percent Correct,” the average correctness for all problems attempted by the student prior to beginning the open response problem (representing a long-term measure of general mathematics ability), and “Prior Sub-Part Performance,” the average correctness across all prior sub-parts of the problem containing the open response question (representing content-specific knowledge).

Additional features were calculated including problem difficulty, the number of prior problem sub-parts, the number of prior problems started by the student, and the number of help requests made on earlier sub-parts, but these were found to either be highly correlated with other measures or not correlated with our outcomes of interest and were therefore omitted from our analyses. All pairwise Spearman correlations of features were calculated, omitting features introducing risks of collinearity.

The inclusion of both long- and short-term performance measures helps to distinguish and control for knowledge of the given mathematical concept as compared to general student ability. While initially hypothesized to be highly correlated, these measures ultimately exhibited a low correlation ($r = 0.014$) and could be used in our analyses without introducing risks of collinearity. Presumably, performance on the prior subparts should be a meaningful predictor of student performance on the open response problem given that they both pertain to the same mathematical concept. If while controlling for this it is found that the student’s prior percent correct is similarly predictive, while not alone causal, such a result would suggest that a teacher may take prior student ability into account when grading the open response.

7.4.2 Factors Related to Student Grades

We use regression analysis to investigate the relationship between the described measure of effort (as measured by time-on-task), teacher-provided grades, and answer- and student-level features using a linear regression model (LM). We additionally include an interaction term of prior sub-part performance \times the

Table 7.3: Linear Regression and Mixed-Effect model coefficients observing assessment score.

	LM($R^2 = 0.133$)		MLM	
	coefficient	std. err	coefficient	std. err
teacher			Variance	Std.Dev.
			0.4523	0.6725
residual			1.5189	1.2324
intercept	0.2803	0.157	1.1648***	0.2226
response category(Words)	-0.3927***	0.033	- 0.3924***	0.0322
prior sub-part performance	1.3684***	0.088	1.3054***	0.0827
words per answer	0.0332***	0.006	0.0393***	0.0053
prior sub-part performance \times words per answer	-0.0185*	0.006	- 0.0189**	0.0057
log of time taken to formulate response	0.0662*	0.020	0.0323	0.0202
prior percent correct	1.275***	0.172	- 0.1497	0.2127

LM: Linear Model, MLM: Mixed Linear Model, Significance: *** < 0.000, **<0.001, * < 0.05

number of words as it was hypothesized that such an interaction would help highlight the relationship between the dependent variable and the combination of skills likely important for open-ended questions.

From the regression results reported in Table 7.3, we found that the model ($R^2 = 0.133$) showed both prior sub-part performance and prior percent correct were reliable and meaningful predictors of student grade. This suggests that both student ability and content knowledge predict student grades. We also found that students with linguistic (word response category) responses correlate with lower scores as compared to responses that contained mathematical terms and expressions. The observed interaction term is found to be statistically reliable, but the low coefficient suggests that the relationship of this term is not very meaningful in comparison to the other more impactful features.

Additionally, we extended the LM by introducing the teacher as a random effect in a mixed-effect linear model (MLM), also reported in Table 3. The grading process, as reported by the teachers in our pilot study, accounts for students' perceived ability. In this model, the prior percent correct measure was no longer a significant predictor of student grade in the MLM suggesting that longer-term performance is not a prominent factor considered by teachers when assessing students. Prior sub-part performance, as a measure of concept knowledge, however, was still a reliable and meaningful predictor of student score.

7.5 Study 3: Potential Impacts of Fatigue on Grading

To address the final research question, we conduct one last analysis to observe how fatigue may affect teacher grades. In other words, we explore whether the ordering in which teachers grade student work leads to any potential risk of unfairness or bias due to implicit sequence or temporal effects. Particularly when a teacher has a large number of students to grade, it may be difficult to grade consistently for all students even when using a rubric. Particularly with the amount of time and attention needed to assess student open response answers, a teacher may find it difficult to give the same amount of attention to the 50th student as they do to the 1st student on a given assignment. Similarly, teachers may grade more strictly or leniently due to unconscious comparisons with previous students (e.g. a mid-grade student response may look better when assessed after a low-grade student response).

To explore this, we use an expanded dataset collected from 2018 to January 2020 to observe the mean and variance of grades over the course of observed grading sessions. This dataset contains 219,189 graded student open responses across 5562 problems from 3847 assignments.

Understanding that teachers may not grade all students for a given problem or assignment in one sitting, we find the order in which teachers graded student answers for each problem on a given day and plot the distribution of grades over this session ordering. Teachers who graded more than 50 students within a single span of time were omitted due to the sparsity of data. We visualize this trend in Figure 7.3 to observe whether fatigue appears to exhibit any temporal effects. If fatigue did affect how teachers grade we would expect to see trends that either affect the mean of teacher scores (rising or falling, on average, as teachers grade more students over time) or the standard deviation of scores (varying more or less as teachers grade).

From figure 7.3, there is little evidence that teachers' grading pattern changes significantly over time both in terms of mean as well as variance (i.e. the width of the distribution does not appear to change significantly, apart from the notable decrease in sample size as the number of problems scored extends above 25-30). While this does not speak to fatigue being an issue for individual teachers or scenarios, it does suggest that fatigue exhibits a low risk in terms of fairness and bias on average across teachers within the system.

7.6 Discussion

Across the three studies presented in this work, we have attempted to gain a better understanding of how teachers approach the assessment of student open-ended work and identify factors that may impact given scores. As it pertains to addressing issues of fairness, the collective results of our analyses provide evidence that, somewhat unsurprisingly, teachers do consider contextual information beyond that pertinent to a given

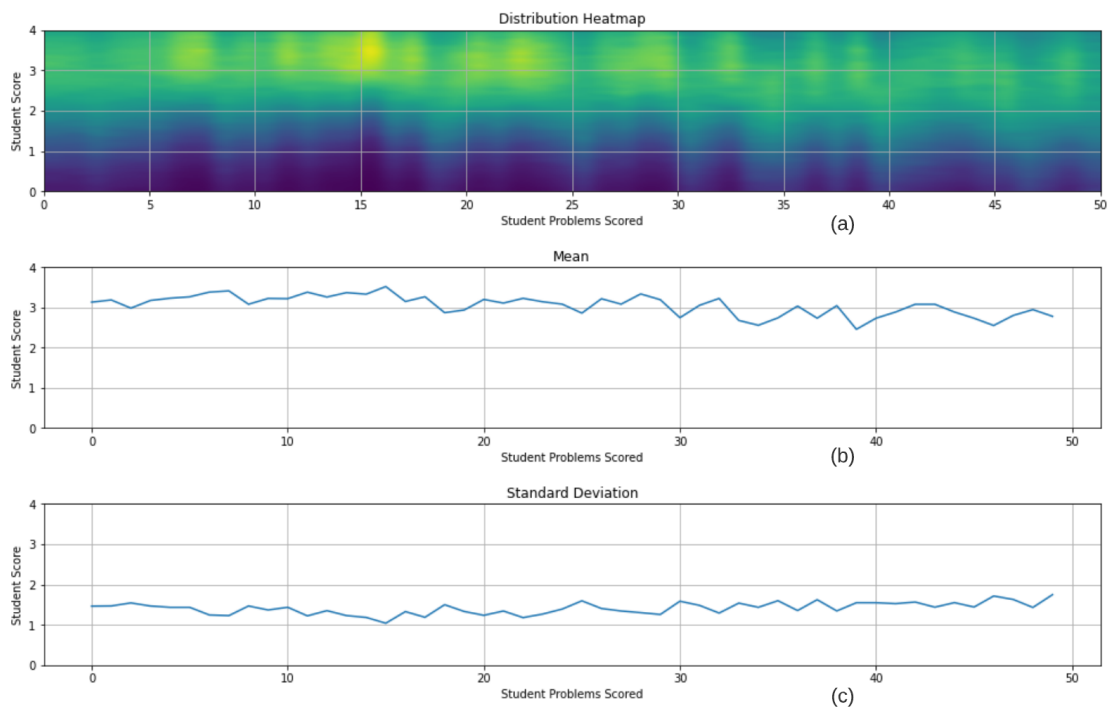


Figure 7.3: The interface for teachers to grade students’ responses on open-ended problems.

answer when assessing student work; this consideration does seem to bias grades in a positive direction, as suggested by the study observing anonymous grading. Qualitative data from our interviews support this, as well as the desire to utilize effort-based measures to even further inform student assessment practices.

From our regression analyses in Study 2, we were able to establish that teachers do not, on average, consider students’ general mathematics ability as much as demonstrated knowledge on the given skill (i.e. in that the measure of concept knowledge was still found to be statistically significant after controlling for teacher effects, while the longer-term outcome was not. This is a little surprising particularly because several teachers from our first study described considering historic student performance when approaching scoring, largely from a motivational perspective for students who typically under-perform compared to their peers.

Finally, though we had initially hypothesized that grading fatigue may be a factor that affects how teachers grade, we found little evidence of this. We found that the mean and standard deviations of teacher scores remained relatively consistent while grading multiple students consecutively.

As we do identify some amount of bias in how teachers grade students, it is important to reiterate that this does not necessarily equate to teachers following unfair or inequitable assessment practices. At the current stage of this work, we are ill-equipped to say whether the measured bias is likely to contribute to impacts on

the types of opportunities that students may receive; we can say from these studies that teachers, on average within the context observed in this paper, seem to be considerate of a range of factors that describe student performance, and are seemingly not largely impacted by obvious negative factors such as fatigue.

7.7 Limitations Future work

As we identified certain biases exhibited in teacher assessment, it is important to explore whether these biases are more prevalent among certain student populations than others to identify deeper risks to student fairness. These populations may refer to protected demographic labels such as race, gender, ethnicity, or other geographic descriptors, but also latent groups of students. Similarly, while no meaningful effect of grading fatigue was observed across all teachers, this may still be an issue in more individualized cases.

In our first study, though we gain some clarity as a result of conducting informal interviews, we are unable to empirically measure if the differences are a result of implicit biases or stereotypes; while teachers identify that they use contextual information to score and motivational factors may factor into their assessment, there may be other explanations that are left unexplored in the current study. Future work could specifically examine this by running a study that explores how teacher perceptions of students (even simulated students) may impact grading practices. Similarly, as certain unmeasured recency or other timing effects may impact how teachers score (e.g. as the teachers re-scored student work from a month prior, and had undoubtedly moved on to new content areas), future work could also replicate this study to randomize when anonymization occurs within the study design.

Among the largest limitations of the current work is the correlational nature of the analyses presented. While we have attempted to pair some of our results with insights from experienced teachers, the causal mechanisms impacting student scores, particularly in reference to those factors external to the student's response itself, could be further explored through additional future studies. It is important to understand what is considered when assessing students so that teachers, education researchers, policymakers, and learning system developers can start to address the questions as to what should be considered and what factors may lead to unnecessary risks to fair student assessment.

7.8 Conclusion

While we are able to identify bias and student-level factors that seemingly correlate with student grades, it is uncertain as to whether such bias is truly negative in regard to fair student assessment. While we did not find any obvious risks to fairness within this presented set of analyses, this work represents a step toward

identifying and mitigating such risks in educational assessment. This work further attempts to emphasize the distinction between bias and fairness, recognizing that the consideration of contextual information can have many positive benefits in terms of student learning, motivation, and engagement. However, it is equally important to pursue further study of these issues to determine whether such benefits implicitly lead to inequitable opportunities.

Chapter 8

EXPLORING THE INFLUENCE OF ANONYMITY AND PRIOR-PERFORMANCE ON TEACHER GRADING BEHAVIOR.

Equity and fairness in assessing student work are paramount for fostering positive learning experiences and ensuring high-quality instruction in classrooms. Previous research has highlighted the susceptibility of teachers to the Halo Effect, particularly concerning factors such as perceived ability and student identity. Building on this body of research, this paper advances our understanding of the Halo Effect's influence on teacher grading behavior through a factorial randomized control trial. The study specifically investigates teachers' susceptibility to student identity (using pseudonyms) and prior performance information, with a focus on educators utilizing computer-based learning platforms for mathematics instruction and evaluation. Notably, the findings reveal no significant impact of either prior performance information or student identity on teacher grading behavior. Teachers consistently displayed similar grading practices across conditions, irrespective of student identity or prior performance information. However, in instances where teachers had graded responses as part of their regular class prior to the study, and had access to the real student identity, there was a variance in the grades when comparing the original grades with those assigned during the study. These findings contribute to our understanding of teacher grading behavior, suggesting the potential existence of *considerate grading behavior*. This behavior entails teachers incorporating student identity into their grading practices to tailor instructional approaches. By examining the complex factors guiding teachers' grading practices, this study contributes to the existing body of research. Crucially, this integration of student identity does not manifest as biases or prejudice but serves as an invaluable tool for acquiring a nuanced understanding of student needs and motivations. The outcome of this study highlights an alternative source of variation in grading schema that can have significant implications for educational research. The findings offer valuable insights into the nuanced ways in which teachers integrate student identity into their instructional strategies.

Proper citation for this chapter is as follows:

Gurung, A. (In preparation). Exploring the Influence of Anonymity and Prior-Performance on Teacher Grading Behavior.

8.1 Introduction

Teachers, in their role as educators, assess students' unique needs and offer personalized instruction and guidance to facilitate learning. However, this evaluation process is naturally prone to personal biases, which can have an impact on their appraisal of students' work. Prior research has explored teacher biases through the lens of the "Halo Effect" [76], a phenomenon characterized by the formulation of judgments based on specific student attributes or characteristics such as likability, appearance, or perceived intelligence. Teachers' perceptions have been found to be sensitive to the Halo Effect in randomized studies exploring the influence of perceived competence [221, 325] and student identity [78, 148]. Even a brief exposure to students' ability, such as watching a short video (~3 minutes) of the student discussing a topic [221] or a short vignette about the student ability [325], could significantly influence the subsequent evaluation of the student's work [221]. Similarly, exploration of student ethnicity revealed a "positive feedback bias" [147] among some teachers where students belonging to racial minority groups, ethnically Black and Hispanic, were graded more leniently and received less critical feedback. Exploration of student gender also revealed that teachers tend to attribute boys with higher ability than girls on partially-correct responses; however, there was no significant difference in grades awarded across genders [78]. While the impact of these two factors, student identity and ability, is widely acknowledged, a comprehensive understanding of how teacher assessment practices are influenced by both sets of information remains unclear. The current work aims to explore this gap in research and explores the sensitivity of teacher grading behavior to student identity and prior performance.

This study intends to draw upon the insights of several prior works that have examined teachers' approaches to grading. Several studies have documented the phenomenon of grade inflation [29, 127, 320, 383], particularly prevalent in higher education settings. Grade inflation can be influenced by several factors, including the direct correlation between assignment performance and grades, which subsequently affects future success, such as college enrollment [383]. Additionally, in higher education, the need to maintain or enhance student enrollment for future classes [127, 29], and the high-pressure situations caused by students' striving to maintain their Grade Point Averages (GPAs) [320], also contribute to this phenomenon. While the phenomenon of grade inflation is well-documented, in a recent study analyzing teacher grading behavior, we observed a different grading pattern among middle school teachers where the priority is to foster learning and less emphasis is put on assignment performance [144]. The study compared the grades the teacher had awarded the students as part of their regular lesson in the prior month with teacher grades when the students were anonymized. We observed a *considerate grading behavior* where teachers demonstrated leniency

when grading lower-performing students and adopted stricter criteria for higher-performing students. This approach enabled the teachers to challenge their high-performing students while providing encouragement to the low-performing students. Our findings demonstrated the ability of teachers to personalize their instruction and feedback by employing their innate judgment and insight into individual students' capabilities. The observation of *considerate grading behavior* further underscores the need to explore the influence of student identity and insight into prior performance on teacher grading behavior.

Building on the findings from our prior study (c.f., [144]), this research uses a randomized controlled trial (RCT) with a 2-by-2 factorial design to better understand teacher grading behavior. Specifically, we explore the influence of student identity and perceived competence on teacher grading practices. To prevent familiarity-based biases, pseudonyms were used instead of actual student names. However, these pseudonyms were chosen to enable teachers to infer the gender and ethnicity of the students. Teachers were provided with average correctness on prior assignments to explore the influence of prior performance. The 2-by-2 factorial design of this study allows us to distinguish between the unique and joint impacts of student identity and prior performance on grading and, importantly, examine how the factors interact and whether these influences vary according to a student's ethnicity and gender. To the best of the author's knowledge, this paper represents the first study to examine the causality of the two factors within the same experimental design. As such, through this paper, we explore the following research questions:

- RQ 1** Does the students' name and inferred information (such as gender, and ethnicity) influence teacher assessment of student responses?
- RQ 2** Does providing insight into students' prior performance influence teacher assessment of student responses?
- RQ 3** Does the ethnicity and gender of the student interact with each other to influence the grading of student responses, and in what ways do these influences manifest?

While this paper investigates the impact of anonymity and insights into prior student performance on teacher grading behavior, it is imperative to note that it does not aim to determine the effectiveness of grades and feedback in improving student performance. Nor does it strive to comprehensively evaluate all the factors that teachers consider in their pursuit of fair and considerate assessment of student work. This work extends previous studies by examining the influence of gender, ethnicity, and prior performance as indicators of student identity and ability.

8.2 Background

Persistent disparities in academic performance across student demographics present a significant societal issue. For instance, despite reduction efforts in the United States, a notable gap remains across ethnicities and genders. African American and Latino students still trail their white peers [71, 375, 122]—especially in STEM fields; and more male students enroll and graduate from higher education programs than their female counterparts [125, 122]. These disparities could carry substantial societal and economic implications, especially considering the increasing emphasis on STEM-related skills and abilities in the job market.

In this section, we examine the factors influencing teacher grading behavior. First, we explore the Halo Effect and a variety of associated factors, including a student’s gender, ethnicity, likability, and perceived aptitude, which can potentially shape a teacher’s perception of student ability and consequentially impact the assessment of their work. Following this, we briefly review existing literature that explores the influence of teacher identity and personal experiences on grading practices. Further, we explore the various approaches that teachers employ when grading student responses. Lastly, to provide a contextual backdrop for our research, we describe the open-ended problems used in ASSISTments [154], a Computer-Based Learning Platform (CBLP).

Furthermore, it’s essential to recognize the inherent link between student identity and the Halo Effect. Traditionally, the influence of student identity on teacher perception has been studied within the framework of the Halo Effect. However, there has been a shift in the educational research community towards prioritizing the exploration of the impact of student identity on teachers. Reflecting on this shift, this paper discusses the influences of the Halo Effect and student identities on teachers separately.

8.2.1 Halo Effect

The Halo Effect is a psychological phenomenon where an individual’s overall impression of a person influences their feelings and thoughts about that person’s character or ability [76]. Essentially, it’s the tendency to believe that if a person is good or desirable in one aspect, they must be good in other aspects too, even if there’s no evidence to support this. Conversely, the Halo Effect can also manifest negatively, commonly known as the Horns Effect ¹, leading to unfavorable assumptions based on one aspect of an individual. Particularly in educational settings, the Halo Effect can substantially impact teachers’ grading behaviors, as their perception of the student is often informed by various student-level attributes. Student-level attributes

¹The Horns Effect is often considered the flip side of the Halo Effect. Instead of associating someone’s positive qualities with an angelic halo, this effect links an individual’s negative traits with devilish horns.

including gender [347], ethnicity [379, 115], “likeability” and attractiveness [200, 64], student names or surnames [203], and perceived ability [221, 220, 325, 20, 19], have been identified as a potential source of bias in grading behavior.

This notion of Halo Effect and its influence on grading biases have been substantiated by prior studies [221, 220, 325] and has been observed to persist across multiple assignments for the same teacher [89]. Furthermore, the prominence of the Halo Effect is particularly notable when teachers assess student writing samples, indicating a susceptibility to bias during subjective assessments [123]. In exploring the nuances of the Halo Effect, some researchers have investigated its persistence over time, while others have delved into teachers’ sensitivity and susceptibility to the effect. Prior studies conducted by Malouff et al. (2013, 2014) [220, 221] and Schmidt et al. (2023) [325] provide compelling insights. In a randomized experiment by Malouff et al. (2014), graders watched a brief video (~3 minutes) of students discussing a topic before grading their work. The findings revealed that teachers graded the same work significantly higher when the same student projected confidence and fluently articulated their ideas in the video, as opposed to instances where they appeared to be less confident and inarticulate. Similarly, Schmidt et al. (2023) [325] reported on a randomized trial where teachers were provided vignettes on student ability (low, medium, and high) on subject A and asked to grade their work. Once completed, the teachers were asked to grade the same student on a second subject, subject B, where the vignettes stated medium ability for all students. However, when grading student work for subject B, teachers in the high ability condition for subject A gave significantly higher grades to their students than the teachers in the other two conditions. These results not only highlight the pervasive influence of the Halo Effect but also underscore the potential of employing insights into prior performance and student ability to augment teaching through personalized instruction and feedback—an aspect that remains relatively unexplored in educational contexts.

8.2.2 Influence of Student Identity

Previous research examining the impact of gender and ethnicity on teacher perceptions of student ability has indicated a tendency for male and white students to be perceived as more proficient in STEM disciplines than their female and black counterparts [370, 201, 293, 354, 147]. Several studies have investigated teacher biases in situations where teachers either had pre-existing familiarity with the students or access to definitive information regarding the students’ gender and ethnicity [70, 71, 114]. Notably, while biases favoring male and white students are often highlighted [77, 78, 147, 149], the implications during the assessment of the students are varied, with some studies reporting Positive Feedback Bias toward Hispanic and African American

students [148, 149], and others finding no significant differences in the assessment of student work across genders [78].

In the following sections, we will investigate the impact of two aspects of student identity, gender, and ethnicity, on teacher perceptions and the assessment of student work. Specifically, we will explore how these aspects of identity can shape a teacher's perception of student work and how such perceptions may, in turn, influence teachers' behavior and students' performance.

8.2.2.1 Student Gender

Researchers have consistently reported on biases among teachers towards attributing higher abilities to boys compared to girls [77, 71, 116, 296]. Intriguingly, such biases are most apparent not with completely correct or incorrect answers but with partially correct ones, where teachers consistently rate male students as more capable [77, 71, 70]. In the context of primary schools, boys are generally perceived as more proficient than girls [70], with girls only being deemed as proficient as their male peers if they also exhibit hardworking and well-behaved traits. This proficiency disparity between boys and girls was corroborated in a *replication study*, affirming the initial findings, conducted 12 years after an initial investigation in 1998-1999 [71]. Further research has probed into teacher perceptions of student aptitude² using the Implicit Association Test (IAT) [138], revealing a male-STEM association among pre-service teachers for grades 1-4 [254], in-service teachers for grades 6-8 [354], and in-service teachers for grade 8 [87].

Although substantial evidence indicates a teacher bias favoring male students in the attribution of ability, particularly within STEM subjects, these findings don't always correspond with performance outcomes. For instance, in a recent randomized study conducted by Copur-Gencturk et al. (2023) [78], 458 teachers were tasked with grading mathematics assignments from students with randomly assigned gender- and ethnicity-specific names. While teachers attributed higher ability to male students in cases of partially correct responses, no discernible bias was observed in the grading process across different genders and ethnicities. Interestingly, research focused on performance measures has sometimes revealed a potential reversal of such biases. For students with similar scores on standardized tests, female students often earn higher GPAs than their male counterparts [116, 201, 293]. While some studies speculate that this phenomenon may be due to the tendency of male students to perform better in more controlled settings [293, 201], these findings are not universal as others have failed to find similar effects [163]. Another perspective suggests the likelihood of

²Though the findings of these studies are not in dispute, it is crucial to recognize that the majority of the studies examining teacher perception of student ability were conducted within European educational contexts. Consequently, while extrapolating their insights, we must remain cognizant of the potential differences in how these phenomena may manifest in Non-European classrooms with respect to both magnitude and significance of the effect size.

female students, unlike male students, fostering stronger interpersonal relationships with their teachers as a potential confound for the disparity in GPA [293] across genders through likability and familiarity.

8.2.2.2 Student Ethnicity

Similar to gender-based biases, biases were observed in teacher perceptions of student ability across ethnicities, wherein a positive association was observed between White students and aptitude in STEM subjects in comparison to Black and Hispanic students [78, 114, 370]. While some researchers have reported on randomized studies that found no discernible bias biases in grading student work across ethnicities despite disparities in perceived ability [77, 78], others have reported on teacher grading behavior manifesting in more interesting ways to accommodate the students' identity. One such example is the "Positive Feedback Bias," a trend identified in previous research [147], where teachers exhibit a tendency to give more positive feedback to student work, with a noticeable bias towards certain ethnicities. Often motivated by a desire to help mitigate educational inequities and societal stereotypes, teachers may unknowingly praise students of certain ethnic backgrounds more frequently and critique them less harshly. For instance, Harber et. al. (2012) [148] reported on a randomized control trial where work ostensibly from Black students received more praise and less criticism, both in-person and in written communication.

While positive feedback bias might seem harmless or even beneficial at first glance, it can have several negative impacts on students. For example, an overabundance of positive feedback may create a false perception of ability, hindering students' ability to identify areas for improvement and growth [65]. Furthermore, a lack of challenging tasks can reduce academic rigor and impede learning and growth [134]. Additionally, prolonged exposure to positively biased feedback can undermine students' trust in their teachers, as they may begin to question whether the feedback they receive is a genuine reflection of their personal achievements or if it is influenced by their racial background [83].

8.2.3 Influence of Teacher Identity and Other Teacher Level Factors

While research into biases has primarily focused on the influence of student identity on teachers, some studies have explored various teacher-level factors that can moderate teachers' grading behaviors. A systematic review of 79 empirical studies pointed to several significant variables [370]. Teacher experience, level of education, and previous exposure to instructing students with disabilities all emerged as salient factors influencing grading practices. Moreover, Wang and Hall (2018)[370] found that a teacher's personal disposition and beliefs about the role of effort in academic success can significantly impact their assessment of student

performance. When delving into the interplay between teacher and student characteristics, the student's race, gender, and disability status were identified as additional elements that can sway teachers' attributions of effort and ability. Highlighting the subtle biases that can infiltrate grading practices, Copur-Gencturk et al. (2023) [78] found evidence of low-performing male students being perceived as having higher mathematical ability compared to their female counterparts. These skewed perceptions can cascade into the nuances of instruction, subtly shifting the tone, body language, and the quality of guidance and encouragement offered to students. While each instruction on its own might seem insignificant, it is entirely plausible that such instruction can have a compounding effect, as student-teacher pairings typically last a year in K-12 settings and at least a semester in higher education.

Additionally, the experiences of biased behavior faced by teachers themselves can also influence their teaching practices. While some individuals may react to such experiences with a diminished sense of self-worth and motivation, others may harness them as a source of empowerment and drive [228]. This aspect of motivation remains relatively unexplored in the context of teacher behavior. Nevertheless, available evidence suggests that teachers' self-image and sense of worth can significantly alter their grading practices. For example, White teachers have been observed to primarily focus on the objective quality and writing mechanics, while ignoring subjective quality, in responses from minority students. This focus on objectivity is potentially an attempt to avoid seeming prejudiced [147, 148, 149]. The multi-layered complexity of teacher grading behavior underscores the need for further research and attention in this field.

8.2.4 Procedural Patterns in Teacher Grading Behavior

Teacher grading behavior often exhibits certain patterns, including a tendency to cluster grades within subgroups of a student cohort. A notable behavior involves teachers deflating grades for high-performing students while inflating grades for low-performing ones [130, 144]. This results in a herding effect where grades are bunched together within subgroups in a class. When applied thoughtfully and with positive intentions, this practice can create a challenging environment for advanced students, while simultaneously offering encouragement to those who struggle. This approach can be particularly valuable in settings where the primary objective is to foster learning and growth, rather than simply evaluate learner ability.

A similar pattern of grade inflation is also observed in higher education. Popularly referenced as "grade inflation," researchers believe this phenomenon is often driven by teachers' desire to avoid confrontations, improve or maintain future student enrollment for the class, and occasional pestering behavior from students under significant pressure to maintain their GPAs [320, 127, 29]. These pressures can stem from various

reasons, including the need to sustain high academic performance to increase eligibility for enrollment into competitive programs, or maintaining eligibility for scholarships and financial aid.

Several potential behavioral factors have been suggested as contributors to grading patterns. One possibility is that as teachers grade a series of responses, they may recalibrate their expectations, leading to a clustering of grades around a class average. Additionally, teachers may tend to grade students in groups based on perceived performance levels, influencing grading behavior. Attention levels could also vary, with teachers possibly being more attentive and meticulous at the beginning of the grading process. Although empirical evidence is scarce, these procedural explanations are supported by teacher testimonies as reported in Rick Wormeli’s book “Fair Isn’t Always Equal” [388]. This underscores the importance of further investigating teacher grading behavior, particularly in scenarios requiring personalized consideration of student needs, to ensure fairness and accuracy while mitigating biases.

8.2.5 Open Ended Problems in Computer Based Learning Platforms

The current study is conducted within the ASSISTments System, a Computer-Based Learning Platform (CBLP) primarily utilized for middle-school mathematics. This platform enables teachers to assign content and automates reporting of student performance, while also providing students with on-demand help and correctness feedback.

Student	Response	Score
Student Name 1	The square root of 9 is 3	▼
Student Name 2	written response	▼
Student Name 3	The area of Q is 45. You have to divide it by the area of P which is 5. The answer is 9. Then, we need to take the square root of 9 and that is our scale factor.	▼

Figure 8.1: Examples of open-ended problems in ASSISTments.

In this paper, we focus on grading open-ended mathematics problems implemented in two distinct formats: (1) as sub-problems within multipart questions, and (2) as standalone problems. Figure 8.1(a) shows an example of the former, while Figure 8.1(b) represents the latter. Notably, a multipart question can contain multiple open-ended sub-problems. After students complete assignments, teachers assess their

responses. Figure 8.1(c) shows the interface teachers use for grading. The ASSISTments system allows teachers to anonymize the process by hiding student names and randomizing response order for impartiality. However, student names is displayed by default.

While some argue anonymization [50, 256] as a potential solution to mitigate biases in performance centered assessments attributable to student identity, it is not without compromise. Particularly in scenarios where fostering learning is prioritized over grades, teachers can adopt a more personalized and considerate approach to assessment [144]—an approach that could potentially minimizes the strength of the relationship between teachers and students [282]. For instance, personalization is both desirable and pragmatic as it enables teachers to fine-tune their feedback and instructional strategies to encourage and support their low performing students while challenging and pushing their high performing students. As teachers have a more comprehensive understanding of student abilities, assignment data can be transformed into an essential component to help enhance the students learning experience. While a personalized approach is ideal for fostering learning, it's not always feasible as gender and ethnic information can inadvertently sway teachers' assessments. The fundamental challenge is to strike a balance: equip the teacher with the necessary context to support students efficaciously while avoiding the influence of factors associated with student identity. A potential avenue worth exploring is leveraging insights into students' prior performance as a tool for informed assessment. By using prior performance data, teachers can gain insights into students' true abilities and needs without the confounding effects of gender and ethnicity information. This approach seeks to strike a delicate balance by empowering teachers with the information they need to tailor their instruction, while at the same time mitigating the risk of conscious or unconscious biases influencing the grading process.

By analyzing teacher grading practices, we aim to explore the intricate ways in which student identity, their prior performance, and the choice to anonymize (or not) can impact the assessment process. This context offers a unique opportunity to investigate the potential effects of combining anonymization with insights into prior performance. Such an approach could promote fair and considerate grading practices while mitigating the influences of biases based on gender and ethnicity.

8.3 Methodology

In this study, we employ a Randomized Controlled Trial (RCT) that utilizes a 2×2 factorial design, focusing on student identity and information on prior performance as the key factors. A specialized tool was developed within the ASSISTments ecosystem to facilitate the investigation of how these factors influence teacher grading behavior. The experimental design enables an in-depth analysis of the individual and com-

Table 8.1: Description of the three types of teacher categories participating in the study.

Category	N	Description
Category 0	7	Teachers who did not have more than a 100 responses to open ended problems from their own students.
Category 1	8	Teachers who had more than a 100 responses from their students and the responses had been graded by the teacher in the past between Jan '22 to June '22.
Category 2	4	Teachers who had more than a 100 responses from their students but the responses had not been graded yet.

bined impacts of student identity and prior performance on grading practices. Subsequent subsections describes the cohort of participating teachers, describes the students' open-response answer dataset used in the experiment, and provides detailed description of the experimental design.

8.3.1 Participants

In this study, we engaged a cohort of teachers who are regular users of ASSISTments [154] for middle school mathematics instruction, utilizing Open Educational Resources (OER) such as the Illustrative Mathematics or EngageNY curricula. Our recruitment process involved sending emails to teachers who have previously shown willingness to participate in research activities with ASSISTments. From those who expressed interest, we selected a diverse group of 19 teachers. Geographically, our participants spanned a broad range, with 18 teachers teaching in 12 different states within the U.S., and one instructing American students in Spain. The cohort comprised 4 male and 15 female teachers.

These teachers were divided into three categories, as illustrated in Table 8.1, based on their usage of open-response problems in ASSISTments from January 2022 to June 2022. Teachers in Category 1 were the most active, having assigned and graded more than a hundred responses to open-ended problems from their students. Category 2 teachers had assigned more than a hundred open-response problems to their students but had not graded the responses. Whereas the teachers in Category 0 had fewer than a hundred open responses from their students during this period. To acknowledge their contribution and incentivize participation, the teachers received financial compensation.

8.3.2 Study Design

Anonymization is the first factor in our experiment. We employed pseudonyms in instances where students were not anonymized. We adopted this approach as teachers might unintentionally utilize student identity to glean additional information beyond the students' gender and ethnicity. Utilizing pseudonyms was particularly important for teachers in Category 0, who were grading responses from students they didn't personally teach. The pseudonyms, along with their associated ethnicity and gender, are presented in Table 8.2. The list contains six pseudonyms for boys and nine for girls³.

Table 8.2: Student pseudonyms used in the experiment with the index. The index represents the order in which the names were arranged in the list and maps to the fixed order of responses within each subsample.

Index	Name	Ethnicity	Gender
1	Jaylen Alston	African American	Boy
2	Jada Jackson	African American	Girl
3	Gabriel Garcia	Hispanic	Boy
4	Antonia Hernandez	Hispanic	Girl
5	Liam Smith	Caucasian	Boy
6	Emma Miller	Caucasian	Girl
7	Peng Chu	Asian	Boy
8	Hitomi Tanaka	Asian	Girl
9	Sanjay Kumar	South Asian	Boy
10	Aastha Valayaputhur	South Asian	Girl
11	Zara Amin	Middle Eastern	Girl
12	Hassan Bilal	Middle Eastern	Boy
13	Brianna Booker	African American	Girl
14	Isabella Lopez	Hispanic	Girl
15	Emily Wilson	Caucasian	Girl

For the second factor, students' average performance on the prior 5 assignments was presented to the teacher when randomized to the condition where they had access to prior performance data. A visual rep-

³The pseudonyms for each ethnicity were generated using online resources. Since the study was conducted with teachers in the United States, the first names were generated using the Social Security and name census websites. The last names were derived using Google search for surnames corresponding to each ethnicity.

resentation of the prior 5 avg assignment performance displayed to the teacher is presented in figure 8.2. In instances where the student had done less than 5 assignments, we displayed a 0 on the prior performance visualization where data was not available.

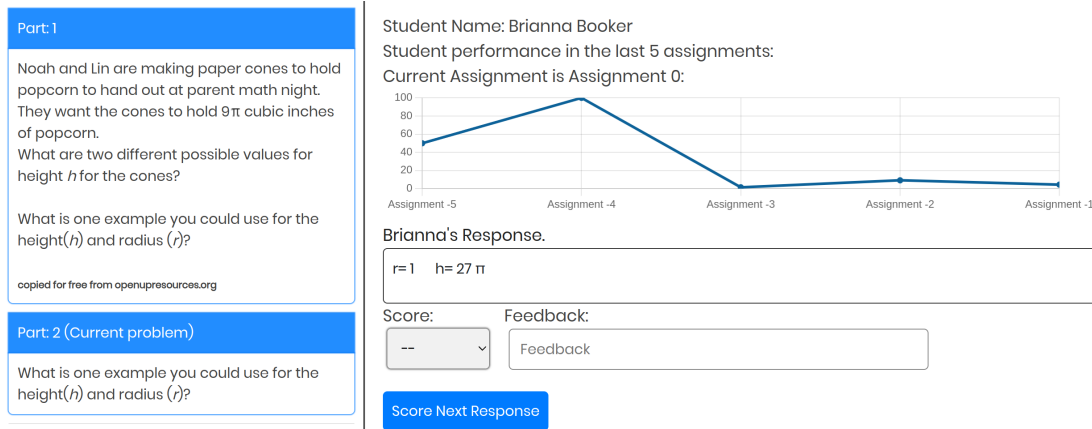


Figure 8.2: A screenshot from the grading tool where a teacher is grading a response from an African American female student Brianna Booker (pseudonym). We display the problem body, prior sub-parts, student's response and average performance on prior 5 assignments.

As previously mentioned, our study entailed the creation of a specialized tool within the ASSISTments ecosystem to investigate how student identity and prior performance impact teacher grading behavior. Initially, we generated a dataset by utilizing real student data from the system. Subsequently, we built a dedicated tool to streamline the process of teachers grading student responses. Below, we outline the logic and approach employed in both the development and deployment of this tool:

- A. Primary Sample:** Each teacher is assigned a random sample of 75 student responses. Teachers in categories 1 and 2 receive responses from their own students, while teachers in category 0 are assigned a random sample of the responses assigned to teachers in category 1—all category 0 teachers were assigned the same random sample with 75 responses.
- B. Sub sampling:** The 75 responses assigned to each teacher are divided into five randomly-selected sub-samples (SS0, SS1, SS2, SS3, SS4), each containing 15 responses. It is important to note that the order of responses within each subsamples remains fixed.
- C. Pseudonyms:** Each student in the randomly generated subsample is given a pseudonyms that is associated with a gender and ethnicity. The list of student names is provided in table 8.2 and the index column is

directly mapped to the fixed order of the students.

D. Batch Creation: As the aim of the experiment is to isolate the influence of student identity and prior performance information on teacher grading behavior, we implemented a round-robin design. This approach involved five iterations (batches), with teachers grading a subset of the 75 responses in each iteration using a 2×2 factorial design. This ensured that by the end of the experiment, teachers had graded the same response four times, once per condition. The pseudo-code below details our method for generating these five batches, each comprising 80 responses. The pseudo-code utilized the third batch as a reference to enhance interpretability:

Step 1 There are 5 batches = [0, 1, 2, 3, 4]

for the third batch: $x = 2$ (the count is 0 indexed)

Step 2 we use modulus (%) to calculate the eligible subsamples: [$x\%5$, $(x+1)\%5$, $(x+2)\%5$, $(x+3)\%5$]

for the third batch, as $x = 2$: SS2, SS3, SS4, and SS0 are the eligible categories

Step 3 The 60 responses from the 4 eligible subsamples are added to the batch.

Step 4 Additional 20 responses that were assigned to other teachers in the eligible subsamples are randomly selected and added to the current batch, making the total 80.

A check is done to ensure that the random sample does not include the responses that are already present in the batch. As category 0 teachers were assigned a random sample of responses assigned to teachers in category 1.

Step 5 the 20 random responses are added to their respective categories.

i.e. problems in SS2 are added to SS2 in the batch; now each subcategory size is ≥ 15

Step 6 The student responses are assigned conditions based on the order of subsample categories:

Subsample Category 1: Anonymized, without prior performance information.

for third batch: Subsample Category 1 is SS2

Subsample Category 2: Non-anonymized (ethnic names shown), without prior performance information.

for third batch: Subsample Category 2 is SS3

Subsample Category 3: Anonymized, with prior performance information.

for third batch: Subsample Category 3 is SS4

Subsample Category 4: Non-anonymized (ethnic names shown), with prior performance information.

for third batch: Subsample Category 4 is SS0

A visual representation of the division in terms of a *2times2* experimental design is presented in figure 8.3.

Step 7 The 80 responses are shuffled and presented in a random order for the teacher to grade and provide feedback on.

E. Washout period: After completing each batch there is a break for 3 days before the teacher is assigned the next batch.

		Prior Performance Information	
		Without Prior Info	With Prior Info
Anonymization	Anonymized	Subsample Category 1 N >= 15	Subsample Category 3 N >= 15
	Not Anonymized	Subsample Category 2 N >= 15	Subsample Category 4 N >= 15

Figure 8.3: A visual representation of the 2 × 2 factorial randomized control trial.

8.3.3 Description of Dataset

To ensure that the teachers were fully prepared and familiar with the study’s tools and expectations, they were provided with an initial training set that included five student responses. Teachers were required to both score and provide written feedback for each response. This requirement fostered more reflective thinking about their scoring as they formulated feedback, thereby enhancing the reliability of the scores. During the study, the teachers were asked to grade 5 batches of student responses as described in the previous section. By the end of the study 300 teacher grades and feedback were recorded with each of the 75 responses being graded once per condition as presented in figure 8.3.

Out of the 19 participants initially enlisted, only 18 teachers actively participated in the study. One teacher belonging to Category 1, despite signing up, did not participate. Comprehensive information regarding the dataset, including relevant descriptors for the participating teachers, is provided in table 8.3. It’s important to highlight that while Category 0 teachers were assigned student responses from teachers in Category 1, some of the responses belonged to the teacher who dropped out of the study. This explains why the sum of problems and student responses in Category 1 and Category 2 does not equate to the overall total in the dataset ⁴.

⁴The data used in this study along with the code is available on GitHub: Exploring Teacher Grading Behavior

Table 8.3: Description of the total problems in the unique teachers, problems and responses per categories.

	Category 0	Category 1	Category 2	Total
Teachers	7	7	4	18
Problems	5	28	27	57
Student Responses	75	525	300	853

Table 8.4: Description of the feedback length (words) and time required (seconds) to provide a score and feedback per score in each condition.

	Score 0		Score 1		Score 2		Score 3		Score 4	
	Length	Time	Length	Time	Length	Time	Length	Time	Length	Time
Mean	13.01	139.63	11.77	607.83	12.26	249.09	10.96	1113.65	4.94	926.47
SD	8.63	1207.99	7.89	9625.59	7.62	3459.50	7.36	22027.76	4.78	24394.28
Min	1.0	4.0	1.0	4.0	1.0	5.0	1.0	4.0	1.0	3.0
Q1	7.0	18.0	6.0	19.0	6.0	22.0	5.0	19.0	2.0	11.0
Median	13.0	27.0	11.77	31.0	11.0	32.0	10.0	31.0	3.0	19.0
Q3	17.0	57.0	16.0	56.0	16.0	55.0	15.0	54.5	7.0	35.0
Max	69.0	304118	53.0	259801.0	49.0	94500.0	51.0	623671.0	63.0	940858.0

8.4 Intra-Rater Reliability

In this section, our objective is to replicate the findings from the prior work [144] regarding the consistency in teacher grading behavior. Our analysis involves a comprehensive examination of the original scores and the responses prior to the experiment, as well as the scores assigned by category 1 teachers to anonymized student responses. We utilize Cohen's Kappa score with linear weights to assess the intra-rater reliability, taking into account the 5-point grading scale. Additionally, we adopt a relaxed Cohen's Kappa approach, considering grades to be the same when teachers are off-by-one between the two conditions. To evaluate the level of agreement, we employ the Cohen's Kappa score ranges [229]: fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and almost perfect (0.81-1.00). Given the focus on intra-rater reliability rather than inter-rater reliability, higher scores indicate more consistent grading practices. Additionally, our analysis extends to investigating the intra-rater reliability in teacher grades across different experimental conditions, with the objective of investigating the consistency in teacher grading behavior under varying circumstances.

A comparison was conducted between the response grades from the first batch of responses when anonymized and the original scores prior to the experiment. The focus on the first batch was intended to mitigate potential

spillover effects, as the study involved teachers grading the same responses in multiple conditions within consecutive batches. In the analysis, two teachers, specifically Teacher 3 and Teacher 5, exhibited notably lower levels of intra-rater reliability and relaxed intra-rater reliability compared to their peers. These findings indicate variations in teacher grades when the responses were anonymized. When a relaxed intra-rater reliability criterion was employed, considering grades off by 1 as consistent, the intra-rater reliability score increased. However, it still remained lower compared to their peers, demonstrating moderate agreement.

Table 8.5: Comparison of Original and Anonymized Scores for intra-rater Reliability of Category 1 Teachers' Response Grades: Replicating Findings from Prior Work [144] on the First Batch of Responses

Teacher	Responses	intra-rater Reliability	Relaxed intra-rater Reliability
	N	Original vs Anon. Score	Original vs Anon. Score
Teacher 1	15	0.72	1.0
Teacher 2	15	0.53	0.81
Teacher 3	15	0.26	0.56
Teacher 4	15	0.67	0.86
Teacher 5	15	0.23	0.56
Teacher 6	15	0.67	0.79
Teacher 7	15	0.61	0.81

As the aim of this paper is to explore the influence of student identity and prior performance on teacher grading behavior we also computed the inter-rater reliability using Cohen's Kappa to compute agreement in scores across conditions for all the teachers by utilizing the anonymized grades without prior performance information as the baseline condition. In general, the intra-rater reliability and relaxed intra-rater reliability across all teachers was high with a nearly perfect score when relaxed as presented in table 8.6.

Table 8.6: Comparison of teacher scores across conditions in the factorial experiment for all teachers.

Condition	intra-rater Reliability	Relaxed intra-rater Reliability
Anon w/o Priors vs. Not Anon w/o Priors	0.74	0.92
Anon w/o Priors vs. Anon w Priors	0.74	0.96
Anon w/o Priors vs. Not Anon w Priors	0.73	0.96

8.5 Main Effects of Anonymization and Prior Performance Information

In this section, we explore the main effects of anonymization and prior performance on teacher grading behavior through our 2×2 experimental design. We employed Python for data preprocessing, while the analysis was carried out using R. Given that our analysis involves multi-level linear regression models, the *lme4* package in R was deemed to be a better fit, as it helped in making our analysis more efficient and streamlined.

8.5.1 Analysis Plan

8.5.1.1 Main Effects of Anonymization

This section describes the method addressing the first research question (**R1**) by analyzing how student identity influences teacher grades, comparing cases where student names are anonymous with cases where names (pseudonyms) are known to the teachers. Utilizing this approach allows us to gauge the average effect of inferred gender and ethnicity on teacher grading practices by considering the scores under anonymous conditions as a baseline. Equation 8.1 specifies the model, wherein both teachers and problems are incorporated as random intercepts. This inclusion accommodates the variability in grading tendencies and scores that can be attributed to either the distinctive disposition of a teacher or characteristics intrinsic to a certain problem, rather than the variables under examination in the experimental setup.

$$response\ grade \sim anonymization + (1|teacher) + (1|problem) \quad (8.1)$$

8.5.1.2 Main Effects of Prior Performance Information

This section elaborates on the approach to explore the second research question (**R2**) by investigating the impact of prior performance information on teacher grading. We do this by contrasting scenarios where teachers were given access to students' prior performance data with those where such information was not provided. For the purpose of this study, a student's average correctness in their five most recent assignments served as an indicator of prior performance. By providing teachers with authentic student performance data, this methodology allows us to gauge, on average, the extent of teacher sensitivity to perceived student abilities in their grading practices, using the grades awarded in the absence of prior performance information as a baseline. Equation 8.2 represents the analytical model, with both the teachers and the problems treated as random intercepts. Such a configuration accounts for grading variations attributable to unique teacher predilections or problem-specific characteristics, rather than the experimental factors under investigation.

$$response\ grade \sim prior\ performance + (1|teacher) + (1|problem) \quad (8.2)$$

8.5.2 Result

The analysis of the main effects from the randomized experiment using multi-level linear regression models as specified in equation 8.1, and 8.2 are summarized in table 8.7. Our analysis revealed no significant difference in student grades between the conditions where student identities were anonymized and those where pseudonyms were used. Though a slight positive effect on grades was observed when teachers had access to student names, this effect was not statistically significant ($\beta = 0.02$, p-value = 0.528, CI=[-0.05, 0.09]). At most, the score on the responses will increase by 0.09 when the students are not anonymized in contrast to when the students were anonymized, as seen by the upper end of the confidence interval.

Table 8.7: Exploring the main effects of the two factors, i.e., the influence of student ethnic names (pseudonyms) prior performance information on teacher grading behavior.

Predictors	grade			grade		
	Estimates	CI	p	Estimates	CI	p
(Intercept)	2.10	1.86 – 2.34	<0.001	2.12	1.88 – 2.36	<0.001
not anonymized (ethnic names)	0.02	-0.05 – 0.09	0.528			
with prior performance info.				-0.02	-0.09 – 0.05	0.597
Random Effects						
σ^2	1.79			1.79		
τ_{00}	0.35	problem id		0.35	problem id	
	0.10	teacher xid		0.10	teacher xid	
ICC	0.20			0.20		
N	18	teacher xid		18	teacher xid	
	57	problem id		57	problem id	
Observations	5400			5400		
Marginal R ² / Conditional R ²	0.000 / 0.201			0.000 / 0.201		

Similarly, the availability of prior performance did not significantly impact student grades. There was a marginal negative effect on grades when teachers had access to students' prior performance, but this too was not significant ($\beta = -0.02$, p-value = 0.597, CI=[-0.09, 0.05]). At most, the score on the responses will decrease by -0.09 when the student's prior performance is provided in contrast to when the prior performance was not available. The small range of the confidence intervals coupled with the relatively small β coefficients suggest that neither of the two primary factors investigated in this study, anonymization and availability of prior performance, had a substantial effect on grading outcomes.

8.6 Influence of Gender and Ethnicity on Teacher Grades

While our analysis reveals that student identity and prior performance information don't significantly affect teacher grading outcomes on average, it's important to recognize that this lack of a main effect does not rule out differences in grades due to the gender and ethnicity of the learners across the 4 conditions, i.e., the 4 cells in the 2×2 design (sub-effects). Bearing this in mind, we delve into the third research question (**R3**) in this section. We examine how student identity and prior performance interact, and explore their individual and combined influences on teacher grading practices. The first cell in our analysis, where students are anonymized and no prior performance data is provided, serves as the baseline. Moreover, when evaluating the influence of the students' gender and ethnicity on their grades, we compare the differences within each subgroup across the 4 conditions.

8.6.1 Sub Effects

This section elaborates on the exploration of sub-effects due to the two factors. Equation 8.3 outlines the model exploring the sub-effects. The teachers and problems are incorporated as random intercepts in the models to accommodate the variability in grading tendencies and scores that can be attributed to either the distinctive disposition of a teacher or characteristics intrinsic to a certain problem rather than the variables under examination in the experimental setup.

$$\text{response grade} \sim \text{anonymization} * \text{prior performance} + (1|\text{teacher}) + (1|\text{problem}) \quad (8.3)$$

The exploration of the sub-effects is presented in table 8.8. Although the differences were not statistically significant, a slight increase was observed in average grades when teachers had access to the student identities without prior performance, and a slight decline in grades was noted when teachers had access to prior performance without student identities. We also estimated an interaction effect indicating that the teachers, on average, scored the students with lower scores when both pieces of information were available or missing as opposed to when one of the two pieces of information was available. Interestingly, there was no average difference between the conditions where teachers had access to both information, student identity and prior performance, and when they did not have access to either—the β coefficient was 0.00⁵.

⁵As the interpretation of the interaction effect can often be erroneous, the formula for the derivation of β coefficient comparing the cell where teachers had access to both student identity and prior performance vs. when they did not have access to either is provided in the appendix section A.4, estimation of sub-effects between the different cells in the 2 × 2 experimental design.

Table 8.8: Exploring the sub effects of the 2×2 factorial design examining the influence of student identity and prior performance information on teacher grading behavior.

<i>Predictors</i>	grade		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	2.11	1.86 – 2.35	<0.001
not anonymized (ethnic names)	0.03	-0.07 – 0.13	0.536
with prior performance info.	-0.01	-0.11 – 0.09	0.840
not anonymized (ethnic names) × with prior performance info.	-0.02	-0.16 – 0.12	0.807
Random Effects			
σ^2	1.79		
τ_{00} problem id	0.35		
τ_{00} teacher xid	0.10		
ICC	0.20		
N teacher xid	18		
N problem id	57		
Observations	5400		
Marginal R^2 / Conditional R^2	0.000 / 0.201		

8.6.2 Sub Effects and Learner Gender

This section explores potential heterogeneity in teacher grades by incorporating students' gender. Equation 8.4 outlines the model investigating these sub-effects and potential variances in grades across learner gender.

$$\begin{aligned}
 \text{response grade} \sim & \text{anonymization} * \text{prior performance} * \text{gender} + \\
 & (1|\text{teacher}) + (1|\text{problem})
 \end{aligned}
 \tag{8.4}$$

Consistent with our prior findings we did not observe any significant effect between learner genders across conditions in the experiment. Although responses assigned to boys in the baseline condition showed a marginally higher average score compared to those assigned to girls, this difference was not statistically significant. Figure 8.4 presents the average scores across conditions for both boys and girls where we did not observe any significant change in student grades across conditions for both genders.

8.6.3 Sub Effects and Learner Ethnicity

This sections explores heterogeneity in teacher grades across learner ethnicity. Equation 8.4 outlines the model exploring these sub-effects and potential variances in grades.

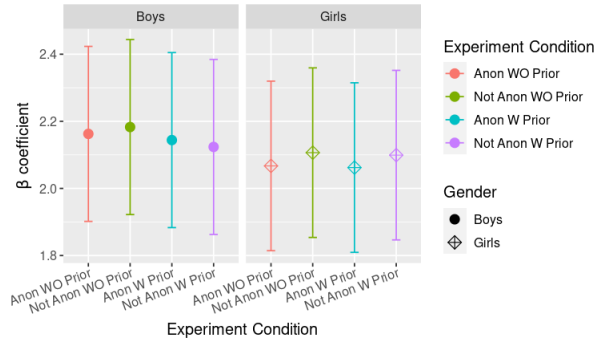


Figure 8.4: A comparison of the difference in average scores between genders as teachers grade students in one of the four conditions in the 2×2 experimental design.

$$\begin{aligned}
 \text{response grade} \sim & \text{anonymization} * \text{prior performance} * \text{ethnicity} + \\
 & (1|\text{teacher}) + (1|\text{problem})
 \end{aligned}
 \tag{8.5}$$

An assessment of the average scores across various ethnicities, using African American students as a reference group, revealed some imbalances in randomization across ethnicities within the baseline condition. Since the responses were randomized, the baseline should have been balanced across subgroups. However, further investigation revealed that the uneven distribution across the baseline conditions was largely due to an imbalance in the random set of responses allocated to teachers in Category 0. As all teachers in Category 0 were assigned the same set of random responses from teachers in Category 1, this relatively imbalanced sample impacted the estimation of the baseline group. An examination of the underlying mechanism that contributed to this imbalance is provided in Appendix A.4.2, specifically in table A.8.

Notably, students identified as Asian ($\beta = -0.33$, p-value= 0.014), Caucasian ($\beta = -0.34$, p-value= 0.005), and Middle Eastern ($\beta = -0.39$, p-value= 0.003) received significantly lower grades. However, differences were not significant for South Asian ($\beta = -0.23$, p-value= 0.081) and Hispanic ($\beta = -0.11$, p-value= 0.343) students when anonymized and without prior performance information. This imbalance suggests that the quality of responses wasn't equally distributed across conditions. Therefore, an average score increase of 0.2 for African American students may not carry the same weight as a similar increase for Caucasian students, and vice versa.

While this imbalance could influence interpretations of results across ethnicities, it doesn't undermine the analysis of average scores within each ethnicity across the four conditions. If significant differences exist within an ethnicity across conditions, it would be necessary to calculate the relative change in scores for a meaningful comparison of how each ethnicity influences teacher grading.

Figure 8.5 illustrates the examination of teacher grades across ethnicities. None of the ethnicities showed

a significant deviation in average scores when compared to their respective baseline category, regardless of whether teachers had access to student identity, prior performance information, or both. The minimal variance can indicate consistency in teachers' grading behavior and suggest a lack of susceptibility to biases based on learner ethnicity.

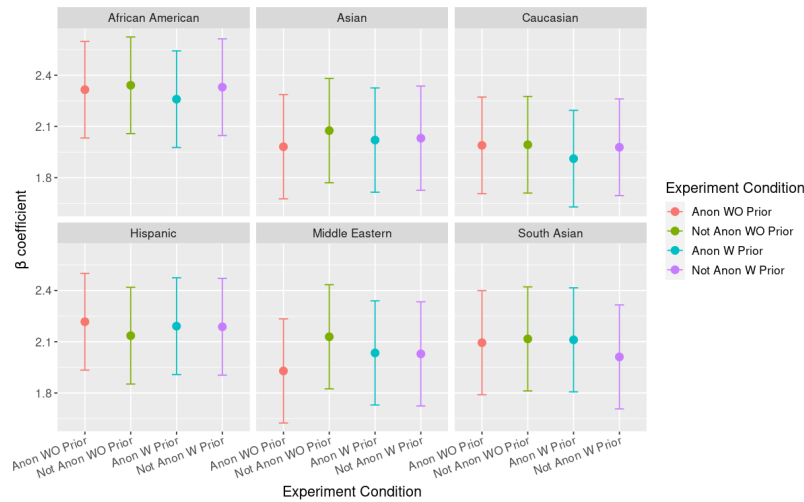


Figure 8.5: A comparison of the difference in average scores across ethnicities as teachers grade students in one of the four conditions in the 2×2 experimental design.

8.6.4 Sub Effects, Learner Gender and Ethnicity

This section uses a four-way interaction between student identity, prior performance, gender and ethnicity to explore heterogeneity in the grades, attributable to the teachers' sensitivity towards information regarding the learners' ethnicity, gender. The model is presented in equation 8.6.

$$\begin{aligned}
 \text{response grade} \sim & \text{anonymization} * \text{prior performance} * \text{gender} * \text{ethnicity} \\
 & + (1|\text{teacher}) + (1|\text{problem})
 \end{aligned}
 \tag{8.6}$$

Introducing the learner gender provided further insight into the discrepancy across learner ethnicity. Similar to the imbalances in the baseline condition across ethnicities in Section 8.6.3, we detected inconsistency in the baseline condition across gender within ethnicity, specifically in categories involving girls; no significant differences were noted for boys when using African American boys as the reference category. For girls, when using African American girls ($\beta = 2.48$, p-value ≤ 0.001) as the reference category, responses assigned to girls in Caucasian ($\beta = -0.59$, p-value ≤ 0.001), Middle Eastern ($\beta = -0.62$, p-value = 0.001), and South Asian ($\beta = -0.37$, p-value = 0.043) categories received significantly lower grades. However, there were no significant differences for Asian ($\beta = -0.32$, p-value = 0.076), and Hispanic ($\beta = -0.19$, p-value = 0.209) cat-

egories. This discrepancy sheds further light on the underlying cause of the significant differences observed across ethnicities in the previous subsection.

Here, it is important to note that for responses categorized as girls, an increase of 0.2 in the average score for African American girls is not necessarily equivalent to the same increase for Caucasian girls, and vice versa. Although this disparity is noteworthy, it does not compromise the within-ethnicity-gender analysis of average scores across the 4 conditions in the experiment. If significant differences emerge across conditions within a specific ethnicity and gender, calculating the relative change in scores will be necessary for a meaningful comparison regarding the influence of ethnicity and gender on teacher grading behavior.

As observed in previous sections evaluating teacher grades across genders and ethnicities independently, figure 8.6 demonstrates that the examination of teacher grades across genders within ethnicities did not reveal any significant differences in average grades attributed to learner ethnicity and gender. Additionally, none of the groups displayed significant deviations in average scores compared to their respective baseline category.

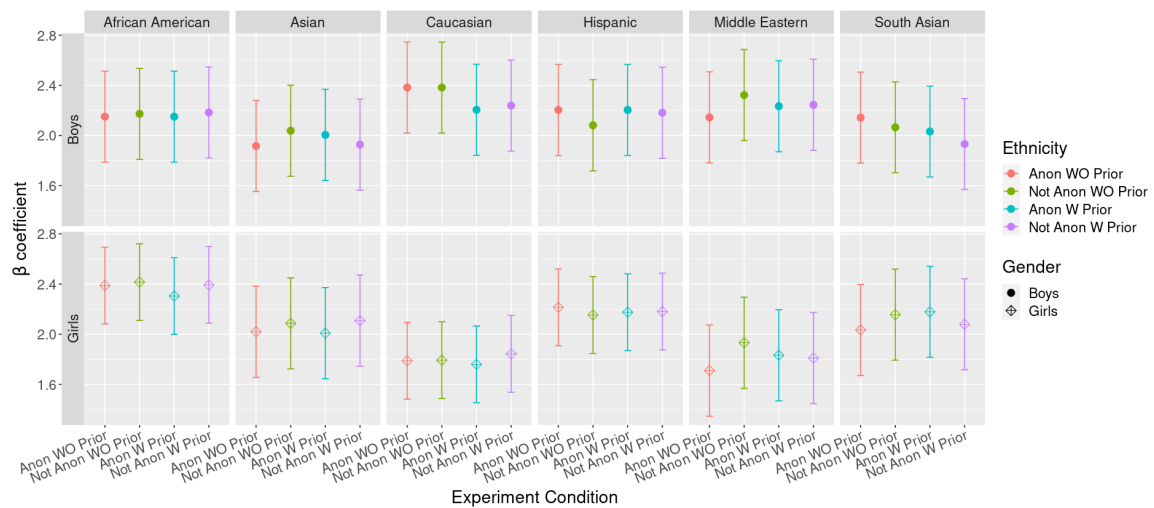


Figure 8.6: A comparison of the difference in average scores across student gender and ethnicities as teachers grade students in one of the four conditions in the 2×2 experimental design.

8.6.5 Results

The sub-effects were explored employing Equation 8.3, as presented in table 8.8. A slight increase in average grades was detected when instructors were privy to student identities, whereas a slight dip in grades was observed when instructors had knowledge of the students' past performance. Intriguingly, the average score remained consistent between the baseline condition (anonymized without prior performance information) and when both sets of information were available to the teachers. It is important to note that the differences

were minor, featuring an increase of 1.42%, a decrease of 0.005%, and no change compared to the reference category, which held an average score of 2.11.

Although these sub-effects did not display statistical significance, this study took additional steps to investigate teacher grading behavior within specific subgroups of student identities. It is critical to examine potential biases at the subgroup level, as biases are often expressed only towards certain sub groups, often remaining hidden in aggregate analyses. By examining the subgroups individually, this study aims to identify grading biases that might be working against or in favor of particular subgroups. Hence, the fluctuation in teacher grades across the four conditions of the experiment was examined, considering scenarios where teachers had knowledge of student identity, past performance, or both. While there were slight irregularities in the randomization, indicating statistically significant differences in the quality of responses assigned to different student identities, these discrepancies did not alter the teachers' assessments across the various conditions within gender and ethnicity. Thus, this study suggests a relatively consistent teacher grading behavior that is not susceptible to inferred gender or ethnic information of the learners. Furthermore, the consistency in the confidence interval across conditions within each subgroup highlight the relatively consistent variance in teacher grades indicating that the susceptibility to Halo Effect and gender/ethnicity to be minimal.

Although certain patterns within subgroups did not display statistical significance, this study undertook additional steps to investigate teacher grading behavior within specific subgroups of student identities, namely gender and ethnicity. Investigating biases at the subgroup level is critical, as biases often manifest only toward certain groups and may remain hidden in aggregate analyses. To this end, the study examined fluctuations in teacher grades across four experimental conditions, considering scenarios where teachers had knowledge of student identity, past performance, or both. While there were slight irregularities in the randomization, suggesting statistically significant differences in the quality of responses assigned to different student identities, these discrepancies did not alter the teachers' assessments within gender and ethnicity subgroups. This indicates a consistency in teacher grading behavior, irrespective of inferred gender or ethnic information. Furthermore, the consistency in the confidence interval, which measures the reliability of the data, across conditions within each subgroup highlights the relatively consistent variance in teacher grades. This points to minimal susceptibility to the Halo Effect based on gender or ethnicity.

8.7 Discussion

This paper employed a factorial design to investigate the influence of student identity and prior performance information on teacher grading behavior. The findings of the study presented mixed results, contra-

dicting previous research that suggested the susceptibility of teacher grades to perceived ability [221, 325] and learner ethnicity [149], while aligning with prior studies that examined teacher grading practices in relation to learner gender [78] when assessing the accuracy of student work. The results from this paper, indicate that the race, gender, or prior performance of students does not significantly impact the grading practices of teachers.

While the exploration of the impact of student identity on teacher grading behavior in this study aligns with prior research, a different approach was taken in the investigation of teachers' susceptibility to perceived student ability (Halo Effect). Earlier studies exploring the Halo Effect utilized videos [220, 221] or vignettes [325], necessitating teachers to absorb information and form perceptions about student ability prior to evaluating their work. In contrast, this study employed a more dynamic methodology to examine teachers' susceptibility to prior performance information by integrating the performance data within the student response. The motivation behind this approach was to assess the feasibility of leveraging prior performance to potentially foster personalized grading behavior, where teacher assessments and feedback can be tailored to the needs of individual students. Such strategies have been suggested as potential means to mitigate identity-based biases while still promoting personalization. By incorporating prior performance information, teachers can more effectively address each student's needs. However, this study did not observe a significant influence of prior performance information on teacher grading behavior.

As outlined in the sub-effects section, this study delved into the heterogeneity of teacher grading behavior by examining the variation in grades based on learner gender and ethnicity. Contrary to previous research, no significant variance in teacher grading behavior was observed across learner ethnicity, indicating the absence of potential Positive Feedback Bias [149]. Similarly, no significant variance was found in teacher scoring of student responses across genders, which aligns with findings from other studies [78]. Taking into account the narrow confidence intervals and their consistency across conditions in the estimation of main- and sub-effects, it appears that teachers demonstrate a high degree of consistency in their grading practices. These non-significant findings contribute valuable insights to the growing body of research investigating the factors that may influence teacher grading behavior. The consistent confidence intervals observed across the four experimental conditions for each subgroup suggest a surprising regularity in teacher grading behavior.

The regularity observed in teacher grading behavior, along with the replication of prior works analyzing intra-rater reliability of teacher grades [144] (as discussed in section 8.4), supports teachers' claims of personalized assessment and instruction tailored to their students' needs. These findings validate the teachers' assertions that variations in grading between original and anonymized scores reflect their intention to person-

alize assessments, drawing on their nuanced understanding of student needs and employing diverse teaching strategies to enhance instructional support and motivate students. The combined findings from this study and previous work [144], where teachers reported adjusting their grading to encourage low-performing students and challenge high-performing students, suggest the existence of *considerate grading behavior* among certain teachers. We encourage researchers investigating fairness and biases in educational settings to incorporate the findings from this paper, thereby enriching their analysis of biases and ensuring accurate recognition and interpretation of *considerate grading behavior* as a positive approach that benefits students.

While the findings of this paper provide valuable insights into the nuanced grading behavior of teachers, it is important to acknowledge several limitations of the study. One limitation is the relatively small sample size of 18 teachers, which may restrict the generalizability of the results to a larger population of educators. Furthermore, the participants were specifically chosen from a group of teachers who regularly use CBLP for teaching mathematics in middle schools, which may introduce a bias towards more technologically proficient and engaged educators. It is important to note that this study differs from previous research as it was conducted digitally, using technology-based assessments instead of the traditional paper and pencil approach. Therefore, caution should be exercised when extrapolating the findings to the broader population of teachers who may have different grading practices.

Another consideration is the specific context of open response problems in mathematics that were used in this study. The nature of these problems may result in relatively low and uniform variance in the quality and correctness of student responses, which can lead to more consistent grading patterns among teachers. It is important to recognize that grading practices may differ in other subject areas or when assessing different types of assignments.

To overcome these limitations, future research should aim to investigate teacher grading behavior in a larger and more diverse sample of educators teaching various subjects across different educational levels. This broader exploration would provide a more comprehensive understanding of grading practices, allowing for a more robust analysis of biases and factors influencing teacher assessment.

8.8 Conclusion

This paper highlights the integrity and conscientiousness in the grading behavior of teachers, particularly in the context of mathematics. It provides evidence against the notion of teacher susceptibility to biases attributable to student gender, ethnicity, or prior performance in teacher grading, and highlights the potential for *considerate grading behavior* based on personalization. The absence of significant variations attests to the

teachers impartial and consistent grading practices. Future research should examine the generalizability of the findings in this paper to larger and more diverse settings, across various subjects and educational levels. It would also be valuable to explore the interplay between *considerate grading behavior* and the contexts in which it emerges, as well as investigating further into how information regarding students' identities and performance may or may not impact grading behaviors in different settings.

Despite the null findings, this study is a significant addition to the ongoing discussion about fairness and equity in education. It emphasizes the importance of continued research and conversation in developing our understanding of the complexities associated with teacher grading behavior.

Part IV

Facilitating Classroom Orchestration and Teaching Augmentation

EXAMINING STUDENT EFFORT ON HELP THROUGH RESPONSE TIME DECOMPOSITION

Many teachers have come to rely on the affordances that computer-based learning platforms offer in regard to aiding in student assessment, supplementing instruction, and providing immediate feedback and help to students as they work through assigned content. Similarly, researchers commonly utilize the large datasets of clickstream logs describing students' interactions with the platform to study learning. For the teachers that use this information to monitor student progress, as well as for researchers, this data provides limited insights into the learning process; this is particularly the case as it pertains to observing and understanding the effort that students are applying to their work. From the perspective of teachers, it is important for them to know which students are attending to and using computer-provided aid and which are taking advantage of the system to complete work without effectively learning the material. In this paper, we conduct a series of analyses based on response time decomposition (RTD) to explore student help-seeking behavior in the context of on-demand hints within a computer-based learning platform with particular focus on examining which students appear to be exhibiting effort to learn while engaging with the system. Our findings are then leveraged to examine how our measure of student effort correlates with later student performance measures.

Proper citation for this chapter is as follows:

Gurung, A., Botelho, A.F., & Heffernan, N.T. (2021). Examining Student Effort on Help Through Response Time Decomposition. In LAK21: The 11th International Learning Analytics and Knowledge Conference (LAK 2021).

9.1 Introduction

Computer-based learning platforms guide students' learning through the implementation of various principles of learning and cognitive sciences. Learning platforms have adopted differing approaches in supporting learners' needs through varying degrees of student- or instructor-paced approaches in determining the content presented to students. In the self-paced paradigm, the systems determine the sequence, and often the difficulty, of content that is presented to the student based on demonstrated performance and mastery of the material; conversely, instructor-paced systems rely on the instructor to determine these assignment parameters. Despite these differences, both of these learning system designs rely on the system to supplement the

instruction and provide additional aid to students as they work; this can simply be done through, for example, immediate correctness feedback, but many systems incorporate more involved instructional aids in the form of hint messages[4, 261, 322], scaffolding problems[328, 385], or other forms of explanations or worked examples. Although the implementation of self-paced and instructor-paced systems often differ, there is a significant overlap in the design principles between the two approaches as both utilize principles of learning sciences and cognitive sciences to enhance learning through these offered supports. These principles have been extensively researched, and various works have explored their effectiveness [263, 313].

Regardless of the learning system's design, there is an underlying assumption that is commonly made regarding student engagement with help provided by the platform. It is presumed that students, when requesting or offered help through the system, are attending to the delivered feedback and using this to learn effectively. While this assumption is likely true for a large population of students, there is certainly evidence that many students take advantage of computer-provided help to work through assignments without effectively learning the material [24, 265]. It is important for students to use help productively, and it is similarly important for instructors to know which students are effectively learning that assigned material.

Our goal in this work is to explore student help-seeking behavior within a computer-based learning platform with a focus on identifying and examining students who are attending to hints they receive through the system. The purpose of this work is to explore this behavior toward the development of a measure of student effort, accounting for systemic differences in the format of help provided (e.g., text-based hint messages or video-based worked examples). In self-paced systems, such a metric could help the system more accurately assess student knowledge and deliver content appropriately [16], or otherwise help instructors monitor and assess student performance more effectively. In either scenario, a measure of student effort, particularly on the help they receive, can help in better understanding the behavior and deploying learning interventions that promote more productive help-seeking strategies.

Using data collected from students interacting with a learning system in real classrooms, we conduct a series of exploratory analyses based on Response Time Decomposition (RTD; c.f., [107, 386, 380]). We further use the findings of these analyses to explore the relationship between identified student help-seeking behavior and later student performance. In this way, this paper addresses the following research questions:

1. Are students using hints appropriately as determined by the amount of time spent on problems?
2. What is the relationship between time spent on hints and later performance?
3. What is the relationship between the time spent on hints and the prior knowledge of a student?

The remainder of this paper is structured as follows. Section 2 describes the related works in the field of Learning Analytics with a focus on student help-seeking behavior. Section 3 explains our theoretical framework that decomposes help usage by users and our hypothesis of the user’s mental model that dictates the actions a user takes after receiving help. Section 4 describes the dataset used in this work and Section 5 breaks down the exploratory analysis conducted to test if the data supports the cogency of our theoretical framework. We use our findings from the exploratory analysis to define user behavior in terms of effort. Section 6 builds on our findings from Section 5 and explores the relationship between effort and other performance metrics. Section 6, 7, and 8 examine our findings and their relevance to research areas in learning analytics to inform future directions.

9.2 Background

Most, if not all, computer-based learning platforms log the actions (clickstream data) of all users interacting with the system. The actions of the students, coupled with measures of performance, are commonly used to generate reports that help teachers monitor student progress. Although these reports provide an overview of the learners’ activity on a given problem set, often in aggregate, the reports provide only limited insight into the learners’ engagement and learning behavior exhibited while working. Efforts in the learning analytics community have helped develop better reports and visualizations that describe several dimensions of student performance, and activity [86, 167]. In this way, developers have attempted to leverage learning analytics research to develop measures that provide finer-grained insights into student learning. Measures of partial credit, for example, help to inform teachers about their students’ knowledge and performance beyond a simple binary correctness measure [374]. Similarly, developing measures of student engagement can better direct teachers’ attention to the students in most need. Researchers have found that the real-time reporting of related measures help teachers spend more time with lower-performing students [167].

The study of help within computer-based learning platforms has similarly led to questions pertaining to the effectiveness of tutor-provided aid within such systems among the learning science and learning analytics communities [15, 156, 34]. In some cases, studies conducted into the role of on-demand help within learning platforms have provided us with valuable insight into help seeking behaviors and various design approaches and principles that can lead to a more effective usage of hints by users [5]; this has been supported, in part, through the study of help-seeking behavior exhibited by learners [4, 358]. Related to this, Researchers have previously studied the use of self-explanation strategies as a method of helping students engage with content [333], while others have explored the format of help delivery through text-based and video-based feedback

[261]. Similarly, researchers have explored the effect of hints versus explanation [133] on student learning, as well as the use of erroneous examples to encourage student engagement with help and learning in general [237]. Finally, there has been other noteworthy research conducted in the field considering how the source or authorship of computer-provided help impacts student learning and engagement [381, 274].

In many cases, these studies have concluded that the effectiveness of help varies greatly and depends on many factors, with perhaps the most prominent of these being the level of student engagement. Regardless of the type of help provided, format, or authorship (e.g., expert-authored versus crowdsourced), these supports cannot help a student who does not attend to and engage with the provided aid. In this way, previous works examining student engagement, or conversely a lack of engagement, are particularly relevant to the study of student help-seeking behavior. Most notably, perhaps, is the large body of work pertaining to the study of students who “game the system” [25, 24, 266, 265]. Commonly referred to simply as “gaming,” this behavior is characterized by students who take advantage of aspects of the system to complete assignments rather than effectively learn the material. In the context of help, students may exhaust available hints [267] or other aids to be given the correct answer or to be given easier questions. Many have theorized and explored aspects that may cause students to disengage, including work pertaining to the study of student affect [109, 46, 237]. Building off these and similar ideas, some researchers have tried to use affect detection to effectively adjust teaching strategies for disengaged behaviour [206], and explore how affect and engagement relate to future student performance [82, 269].

It is clearly important to promote engagement among students and to similarly promote positive help-seeking strategies, but it is also the case that engagement and persistence is not always productive. The example of “wheel spinning” behavior (c.f., [35]), for example, illustrates the negative aspects of persistence. Wheel spinning is defined as a student’s struggle to master a given skill despite being given multiple practice opportunities; practically speaking, wheel spinning as been previously defined as a student being unable to demonstrate understanding of a concept by answering three consecutive questions correctly by the tenth item on a mastery learning assignment [35]. In light of wheel spinning behavior, and in consideration of the many works referenced in this section, it is important to identify students who are truly struggling and where the computer-provided help is failing to aid them. Toward this, it is the goal of this work to develop a measure of student effort as defined by engagement and attentiveness to assigned work. We seek to distinguish students who are applying effort from those who may appear to be exhibiting wheel spinning, but are, in actuality, not “spinning their wheels” in the context of computer-provided hints. This paper focuses its attention to the sub-action level, observing variations in time between requested help actions within a learning platform to

examine these aspects of student learning and help-seeking behavior.

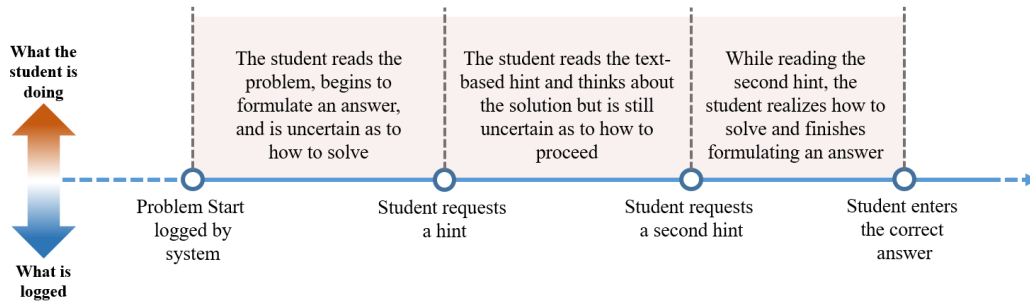


Figure 9.1: Visual representation of the student behaviour for a user interacting with a Computer-based learning platform.

9.3 Theoretical Framework behind decomposition of help usage

While students work through assigned problems, regardless of the learning platform, there is a subtle disconnect between what is being logged and the learning processes taking place. It is certainly the case that actions logged by a learning system provide evidence to latent learning constructs (e.g., knowledge[80, 270]), as the actions were taken by a student and aspects of those actions (i.e., correctness) provide evidence of underlying cognitive and behavioral processes. However, these actions are not direct measurements of these latent attributes and must be viewed in conjunction with expectations as to what occurs between actions logged in the system to gain better insight into processes of learning.

Consider, for instance, the example illustrated in Figure 9.1. In this example, a student begins a problem in a learning system and is able to ultimately reach the correct solution after receiving help. From the perspective of the system, what is logged is just four actions: the start of a problem followed by two help requests, and an attempt to answer with the correct solution. However, in that example, the actions themselves are not able to represent shifts in an activity that occurred external to the system. It is not, for example, able to capture when the student finished reading the question and began to think through how to formulate a response. We can hypothesize that the student was perhaps confused or lacked the knowledge to solve the problem in that the student requested a hint, but there is a large degree of uncertainty as to what the true reasoning for the action was in addition to the sequence of actions, behaviors, and thoughts that occurred external to the system between the start of the problem and the help request.

In order to measure these actions and behaviors, there are several approaches that can be explored. First, the use of additional sensors (such as video) or human observers can help record activity that occurs outside the learning system; such methods have previously been applied to study mind wandering [234] and student

affect [100, 45], for example. These methods, however, can be potentially intrusive, expensive, and difficult to implement in classroom settings due to other ethical and privacy concerns. Another method is that of self-reports. By asking a student to reflect on their thought processes, we may be able to gain insights into aspects of the student’s approach to problem-solving that was missed by the system. This method, however, can be potentially disruptive depending on when it is asked, or unreliable if the student is not able or not willing to articulate their approach with precision. The last method is the examination and analysis of data to make inferences of student activity based on the evidence provided through those actions that are logged and the time between them. While not as definitive as the other methods, as it is more difficult to externally validate many of the inferences made, this method can be applied post-hoc to large amounts of data without facing the concerns exhibited by the other two methods.

Given the actions logged by the system, coupled with the time between those actions, we hypothesize that we can gain insight into the productivity of student usage of help by decomposing the time spent after requesting help in a learning system. In the example illustrated in Figure 9.1, the student read through each requested hint and took the time to think through the new information as it related to formulating the correct solution; it is theorized that such students who are attending to the help would spend more time after the hint and would be more likely to answer the following, related problem correctly than a student who does not exhibit the same effort. By observing the response times in conjunction with the following actions, we hope to gain this measure of effort, even if we are unable to specifically identify the specific latent processes exhibited beyond this valence metric.

9.4 Description of DataSet

For our exploratory analyses, we collected a dataset ¹ by randomly sampling 20,000 student-assignment interaction logs from ASSISTments [154] from the 2018-2019 and first half of the 2019-2020 school years (i.e., before the shift to remote learning in response to the COVID-19 pandemic). ASSISTments is a computer-based learning platform that allows teachers to assign content (primarily in the domain of middle-school mathematics) and monitors student progress, while supplying students with immediate correctness feedback and, on many problems, computer-provided help in the form of on-demand hints and scaffolding. Teachers are able to assign several types of assignments including a “complete all” that requires students to complete all assigned problems (similar to traditional paper-and-pencil assignments with the added benefit of computer supports), as well as “skill builder” assignments, which instead are mastery-based; skill builder assignments

¹The data and code used in this work are made publicly available at <http://tiny.cc/LAK21-28>

require students to demonstrate an understanding of the material by answering 3 consecutive questions correctly on the first attempt without the use of computer-provided aid. The data used in this paper observes both types of assignments but is primarily composed of skill builder work.

While working through assigned problems in ASSISTments, students are able to make multiple attempts to answer as well as receive aid by requesting help in the form of hints (available on many problems in the form of either text- or video-based messages and examples), or scaffolding questions that help break the problem into smaller steps. Problems may contain multiple hints which may be requested by the student, where, in all cases, the final “bottom-out” hint provides the student with the answer. Students are not able to move on to the next problem without eventually providing the correct answer.

The dataset contains the action logs from students who started work on the randomly-sampled assignments. Overall, the dataset contains 644,095 action logs from distinct 14,824 students working on problems across 6,569 problem sets that have a total of unique 36,441 problems. The difference between the total users and assignment logs indicates that we have records for users who did more than one assignment on the platform. The purpose of randomly sampling student-assignment interactions in this way was an attempt to create a sizeable dataset that is not based on a particular subset of content or groups of students; the selection of 20,000 such logs was an arbitrary decision, but we argue is sufficient to conduct the analyses and make impactful claims regarding the observed behaviors of students therein.

In our context, an action is logged every time a user interacts with the system. The system logs actions, for example, when the users start the assignment, start working on a problem, make an attempt, ask for help (as hint, explanation, or request for the correct answer), complete a problem and complete the assignment, among others (there are many system-level actions that can be taken describing a student ending a session and resuming, for example). Each action is accompanied by a timestamp to indicate when each action was taken by the user on the system. The dataset has a unique identifier for each individual user and each assignment as well as other descriptives including, for instance, the start and end time for each assignment. The dataset also has unique identifiers to represent the problem set and the problems the users are working on.

As we are interested in decomposing the amount of time a user takes between actions, we explore the data in regard to action pairs representing sequences of recorded actions; as exemplified in Figure 9.1, it is the goal of this work to take a step toward identifying processes that occur between actions and intend to use the observed time between actions as a means of addressing this goal. We first combined all the actions into pairs, denoted throughout this paper in the form “(first action, second action)” where these represent two consecutive actions taken within the session (i.e., we do not consider an action pair where the student logged

Table 9.1: filtered Action Pairs of students who asked for a hint

Action pairs	N
(Hint Request, Attempt)	808
(Hint Request, Hint Request)	414

out and resumed before continuing). Action pairs help us calculate the amount of time, in seconds, a user took after an action before taking the next action. While exploring the data, we discovered that the time a user took between first and second action ranged from close to 0 seconds to, in a small number of cases, more than an hour; as such, we applied a natural log-transform to the student response time to observe trends and relationships using the measure as an approximate-normal distribution.

9.4.1 Action pairs considered

As it is our goal to decompose student response time in regard to help-seeking behavior, we filtered the action pairs to include only those involving student help requests from the system. This work excludes the observance of scaffolding requests and instead focuses on hints within the system; as scaffolding problems may offer hints themselves, a deeper exploration of this type of aid is more complex and is planned as part of future work. Particularly, there are two notable types of hint requests that existed within the dataset: hints and explanations. The system defines these as separate forms of help, with hints often occurring in a series (i.e., there may be multiple hints), while explanations are singular and give the answer to the student following instruction or a worked example. We found, in our dataset, there were very few samples containing explanations, and fewer samples where the student actually requested such an explanation. As such, we further limited our analyses to explore only hint requests made within the system. We also excluded requests for the last hint in the sequence, referred to as the bottom-out hint, as this gives away the answer; we do not expect students to attend to the given answer in the same manner as a more-instructional hint, and therefore limit the scope of this work to focus specifically on non-answer-giving hints. Given this filtering to examine only hints, we will refer to help within the analyses described in this paper as “hints” to avoid conflating results with potential differences that may be examined in future works regarding other forms of help.

From this, we observe two primary types of action pairs, distinguished by the subsequent action taken after requesting a hint in the system. The intuition behind this is that students likely take additional time to formulate an answer when the subsequent action is an attempt as opposed to another hint request, or

otherwise the response time is likely to incorporate different processes that lead to the different subsequent action. Namely, these action pairs are:

- (Hint Request, Attempt): The action pair (Hint Request, Attempt) represents all the instances when the user asked for a hint from the system, and the next action the user took after getting the hint was to attempt to answer the problem.
- (Hint Request, Hint Request): The action pair (Hint Request, Hint Request) represents all the instances when the user asked for a hint from the system, and the next action the user took was to ask for the next hint.

In order to explore the theoretical framework behind decomposing help usage, we look at the instances when the user asked for a hint or multiple hints within the first 4 actions of working on a problem for both the action pairs. The action pair time represents the amount of time the user spent analyzing the hint before taking the second action in the action pair. We then z-scored the action pair time taken (again, represented as log-time) for each action pair and filtered the records with a value outside of the range $(-3, 3)$; this filtering step is an attempt to remove very large outliers that may influence our results in unpredictable ways. The final resulting number of action pairs used in our analyses are shown in Table 9.1.

9.5 Exploratory Analyses

In this section, we discuss the response time decomposition exploratory analyses conducted in examining student hint usage. As part of this, we examine not only differences in response time, but also explore potential systemic explanations for any differences observed (e.g., the format and length of hints requested). We used python for our analysis and the plots were generated using the Seaborn data visualization library. The y-axis in the charts of this section are the Kernel Density Estimation of the Gaussian distribution.

9.5.1 Analyzing action pairs

First, we observe student response time comparing the second action taken in regard to the first action that students take on the given problem. In other words, we hypothesize that students may use help differently depending on if they felt confident enough to attempt the problem before requesting a hint as opposed to requesting a hint as the first action on the given problem. As such, we observe first the time taken across all first actions and compare this to only the students who request a hint as the first action on the given problem.

9.5.1.1 Examining students across all first actions

We analyzed the two sets of action pairs by plotting the log-transformed distribution of the time taken across students exhibiting each of the action pairs. We found that the distribution of the (Hint Request, Hint Request) action pair to be distinctly bimodal in nature whereas the (Hint Request, Attempt) appeared to be closer to a unimodal distribution. Figure 9.2 shows the overlaid distribution of both action pairs.

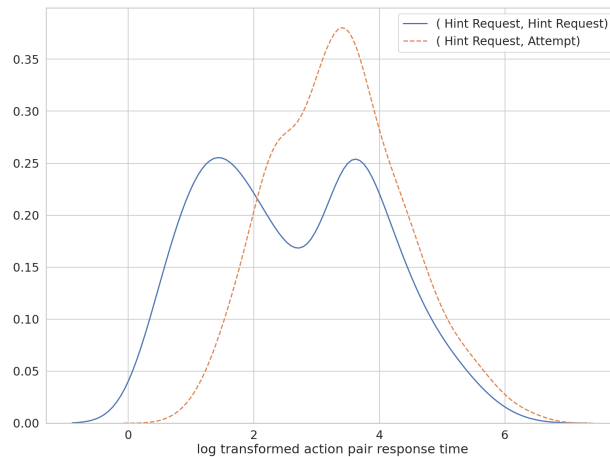


Figure 9.2: distribution curve of (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs using natural log-transformed values of time taken for each action pair

The distribution illustrated in Figure 9.2 suggests that the users who ask for a hint and make an attempt to answer the question are similar to users spending more time on hints; we hypothesize that these students may be those who spend more time attempting to understand and appropriate the information given by a hint before taking a second action. The alignment between the students spending more time on hints with those students who attempt an answer following a help request suggests that these students may be related in their usage of the hint; of course this claim cannot be verified from this plot alone, but does align with our theory that students who spend more time on help may be using that time productively to remedy gaps in knowledge. This also helps us intuit that users in the first half of the (Hint Request, Hint Request) action pair distribution (i.e., the left “peak” of the bimodal distribution) may not be devoting the same attention to the hint as those students spending more time; the cause of this is unclear, however, as it could suggest that these students are not reading or attending to the hint, but it could also suggest that these students are able to recognize that the hint is not helpful early and request a second hint in search of the information they need.

9.5.1.2 Examining students who request a hint first

In order to further refine our analysis, we also analyzed the response time for users whose first action after reading a problem was to ask for a hint. Figure 9.3 shows the normal distribution of both action pairs; we used the natural log-transformed values of the two pairs as that allows us to compare the two distributions. It is important to note that there are many similarities found between this and Figure 9.2, with the largest differences being seen in the shape of the (Hint Request, Attempt) distribution; we use the description of “differences” with hesitation here as there were very few meaningful differences between the two distributions.

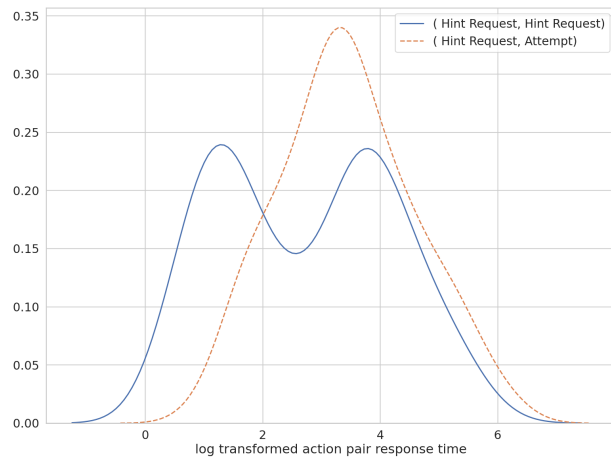


Figure 9.3: distribution curve of (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs, when their first action was asking for hint after reading the problem, using natural log-transformed values of time taken for each action pair

Again, while subtle, the distributions depicted in Figure 9.3 show some variations. The (Hint Request, Attempt) action pairs distribution, for example, appears to be slightly smoother than was observed in Figure 9.2. This is rather unsurprising as we would expect observing the distribution of this subset of students would result in a smoother distribution, however, the smoothing shifts the mean of this distribution in favor of longer response times. This suggests that students who ask for a hint as the first action and make an attempt to answer as the second action, such students are spending more time on the requested hint. No such trend is observed for the students who are requesting multiple hints.

9.5.2 Examining Potential Systemic Causes

In order to better understand our observations in regards to the response time during hint requests, we explore the existence of any potential systemic causes driving user behavior in both the (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs. For the (Hint Request, Attempt) action pair, we also explored

if the correctness/incorrectness of the user's subsequent attempt impacted the nature of the action pair's time distribution.

9.5.2.1 Video vs Text

The system can provide hints to a user as a text or video. We wanted to explore if the format of the hint influenced the amount of action pair time observed, particularly examining whether this formatting could explain the bimodal distributions observed in the previous plots. Figure 9.4 shows another seemingly-bimodal distribution of the (Hint Request, Hint Request) action pair and the shape of the distribution when we only take text hints vs video hints; we used the log-transformed values of the two pairs as that allows us to compare the distributions as was conducted in the previous analysis.



Figure 9.4: There are too few instances of video hint for us to draw a conclusion but the data does seem to indicate that the type of hint does not influence the action pair response time

Figure 9.5 shows the normal distribution of (Hint Request, Attempt) action pair and the nature of the distribution when we only take text hints vs video hints; we used the two pairs' natural log-transformed values to compare the distributions.

9.5.2.2 Correct Attempt vs Incorrect Attempt

In observing action pairs containing an attempt as the second action, we further examined if there were any meaningful differences in response time when the attempt was assessed to be correct as opposed to incorrect. Figure 9.6 shows the distribution of (Hint Request, Attempt) action pairs for these attempts. It can be seen in this figure that students tended to spend less time on incorrect attempts, but does not exhibit a large, meaningful difference; the distributions follow a nearly-unimodal shape despite the observed trend.

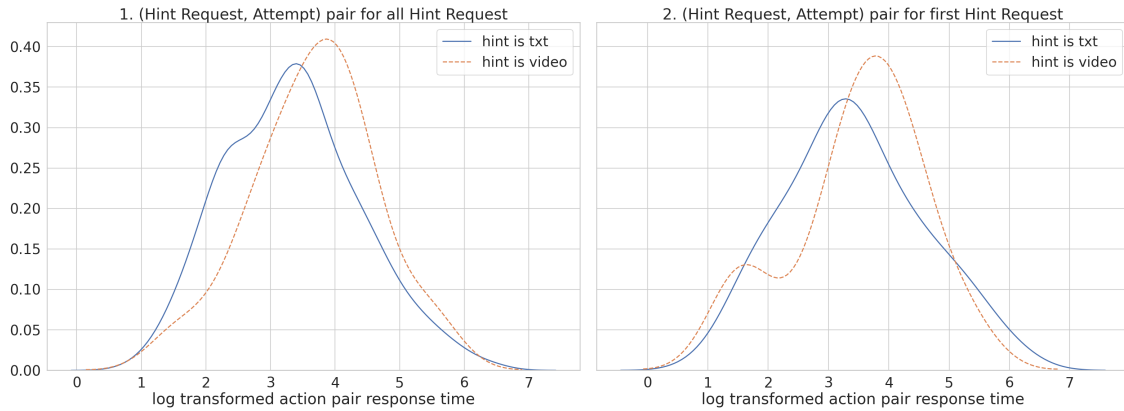


Figure 9.5: There are too few instances of video hint for us to draw a conclusion but the data does seem to indicate that the type of hint does not influence the action pair response time

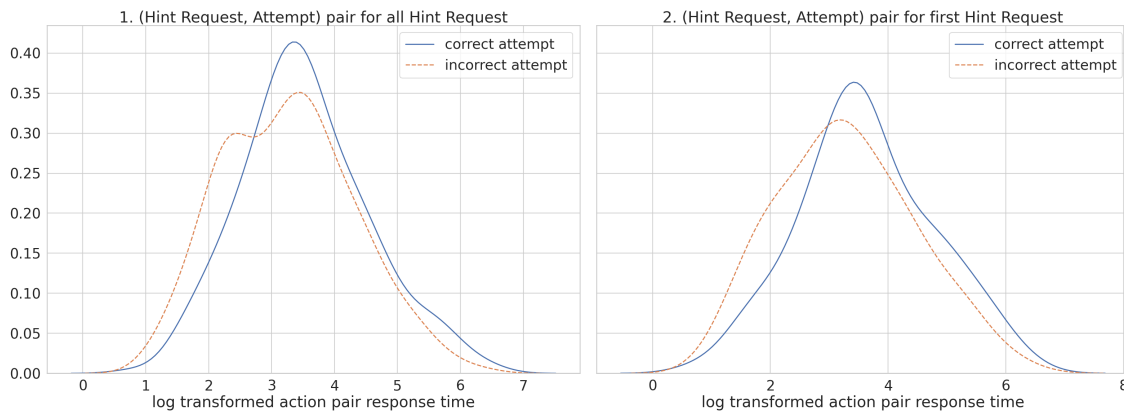


Figure 9.6: The amount of time a user spends after getting a hint is the same for students who made a correct or an incorrect attempt

9.5.2.3 Other Explored Systemic Explanations

In addition to the systemic explanations explored above, we additionally examined the content of hints to observe whether the length and inclusion of visual components such as tables and mathematical formulae explained some of the differences in response time observed in the previous plots. These observations are summarized below; plots are not included for these due to spacing constraints.

1. Length of Textual Hints: We analyzed the amount of time a user spent trying to understand a hint based on the length of the hint. The hints were divided into 4 quartiles based on the number of words per hint. We found users investing more time to understand hint when they were given a shorter hint i.e., hints with less than 18 words. We did not find a difference in the correctness of subsequent user attempts based on the length of the hints. While the length of hint did correlate with the amount of time spent

after the request, the same bimodal distribution emerged as before, suggesting that the length of hints did not explain away this observed difference.

2. Tables and Formulae: We found some hints contained visual content such as tables and formulae. Performing a similar visual analysis comparing the response time for such cases, the resulting distributions did suggest that the inclusion of such content is correlated with higher observed response times, but, similar to the number of words in the hint, did not explain the previously-observed bimodal distributions. It is difficult to make stronger claims in regard to this finding, however, as the presence of tables and formulae in hints was too sparse in the data.

9.6 Examining Student Effort

9.6.1 *Defining Effort*

Our findings from the exploratory analysis, in the previous section, of the response time decomposition of users upon receiving help(hint) goes to support our theoretical model of user behavior. As the user response distribution for (Hint Request, Hint Request) action pair is bimodal in nature and the (Hint Request, Attempt) action pair distribution overlaps with the second peak of the bimodal distribution we use the information to formulate our definition of user exhibiting effort upon receiving help from a computer-based learning platform. In our theoretical model, we hypothesize that the amount of time a student spends on a problem trying to solve the problem is influenced by their understanding of the problem and the underlying concept the problem is trying to address. The amount of time they spend trying to understand the hint provided by the system is influenced by their understanding of the core idea behind the problem and the soundness of their mental model they formulated in order to solve the problem. A student sincerely trying to solve the problem would put in time understanding the hint, recalibrating their mental model to solve the problem, and decide if they have the answer or they need further help. Using the evidence from our analysis we hypothesize that the students in the first hump of the distribution for (Hint Request, Hint Request) action pair are not putting in the effort to understand the hint hence we define those users as exhibiting “*low-effort*” on the problem, the students in the second hump, we believe, put in the effort to understand the hint and tried to formulate an answer using the hint hence we define those users as exhibiting “*high-effort*” on the problem.

9.6.2 *Modeling Student Effort*

The students exhibiting high-effort on both action pairs (Hint Request, Hint Request) and (Hint Request, Attempt) overlap on their time distribution for high-effort behavior; we merge our two action pairs into a sin-

Table 9.2: the mean(μ) and standard deviation(σ) for the high and low effort clusters using Gaussian Mixture Modelling

	mean(μ)	standard deviation(σ)
Low-effort	1.7	0.757
High-effort	3.9	0.909

gle action pair (Hint Request, Action). As our primary interest is on decomposing user response to help and the amount of time a user spends unpacking the hint. As this distribution is bimodal in nature we apply Gaussian Mixture Models(GMM) to calculate the likelihood of the time spent by the student, understanding the hint, is part of the distribution of high-effort users, and the likelihood that the user is part of the distribution of low-effort users. GMM are a probabilistic model of representing a normally distributed subpopulation within an overall population. GMM is an unsupervised learning algorithm that uses Expectation Maximization to cluster the observations in a population into a subpopulation using probabilistic estimation that it is part of a subpopulation within the overall population. We clustered the bimodal distribution into two clusters using GMM; Table 9.2 shows the mean(μ) and the standard deviation(σ) of the two clusters.

We now use the mean(μ) and the standard deviation(σ) from the two clusters to calculate the area under curve for every response time if it were part of the low-effort distribution and the high-effort distribution. This provides us with insight into where the response time falls in the low effort distribution and high effort distribution if it were a user exhibiting low or high effort respectively. We realized that there were three major regions in the distribution where a user response time can fall. For the instances where the area under curve is less than 50 percent for low effort, we label them as low effort and for the instances where the area under curve is larger than 50 percent for high effort we label them as high effort; however, for the instance that do not meet these requirements we can intuit the effort exhibited by the user but we cannot definitively say if they are exhibiting high or low effort so we did not label them.

9.6.3 Exploring the Relationship Between Effort and Performance Metrics

To explore the relationship between our measure of student effort and later performance metrics, we paired the action-level data used in previous analyses with both prior and later student performance measures. These additional measures include assignment completion, wheel spinning in the assignment (as defined by [35]), next problem correctness, prior percent correct (i.e. the percent of problems answered correctly by the student prior to each observed problem), and prior completion rates. We wanted to investigate if the

Table 9.3: Logistic Regression analysis exploring the relationship between effort and next problem correctness while controlling for prior percent correct ($R^2 = 0.048$)

	coefficient	std. err	conf. interval	p-value
intercept	-1.7747	0.360	[-2.481, -1.069]	0.000
High effort	-0.2652	0.271	[-0.797, 0.267]	0.328
Low effort	-0.7053	0.343	[-1.378, -0.033]	0.040
Prior percent correct	2.2975	0.615	[1.091, 3.504]	0.000

students exhibiting effort perform better in the immediate next problem, if they are more likely to complete the assignment, and if they are more likely to exhibit wheel spinning during the assignment.

We used regression analyses to investigate the relationship between student effort and each of these outcome measures while controlling for prior completion rate, prior percent correct, and prior completion rate respectively. The observed models and results of our regression analysis are observed in Tables 9.3, 9.4, and 9.5, and are discussed further in the next section.

9.7 Results

We trained a logistic regression to explore the relationship between effort and next problem correctness while controlling for prior percent correct; it is important to highlight, as this is a logistic regression, that the coefficients are reported in log-odds units and should therefore be interpreted in terms of their magnitude rather than in terms of standard deviations or percents as is commonly afforded by linear regression models. We found that the model ($R^2 = 0.048$) showed that low effort behaviour, $\beta = -0.7053$, $p=0.4$, was a significant predictor of next-problem correctness. This suggests that students exhibiting low effort are more likely to answer the next problem incorrectly. The same cannot be said for the students who are exhibiting high effort. The regression analysis is reported in Table 9.3. It is also important to note that the r-squared of the model is relatively low, which, while it does not detract from our findings, suggests that there are other larger factors that we did not account for that explain the dependent variable (e.g., likely other skill- or content-based factors).

We also examined the relationship between effort and wheel spinning behavior while controlling for prior completion. We found the model ($R^2 = 0.091$) found that low effort behavior, $\beta = 1.0741$, $p < 0.001$, was a significant predictor of wheel-spinning behavior. The analysis found that high effort behavior, $\beta=-0.5815$, $p=0.053$ was a strong indicator of wheel spinning behavior. This indicates that the students who are exhibiting

Table 9.4: Logistic Regression analysis exploring the relationship between effort and wheel spinning while controlling for prior completion ($R^2 = 0.091$)

	coefficient	std. err	conf. interval	p-value
Intercept	0.3809	0.387	[-0.378, 1.139]	0.325
High effort	-0.5815	0.301	[-1.171, 0.008]	0.053
Low effort	1.0741	0.294	[0.497, 1.651]	0.000
Prior completion	-1.8236	0.502	[-2.808, -0.840]	0.000

Table 9.5: Logistic Regression analysis exploring the relationship between effort and assignment completion while controlling for prior completion ($R^2 = 0.104$)

	coefficient	std. err	conf. interval	p-value
Intercept	-3.3584	0.484	[-4.307, -2.410]	0.000
High effort	0.3614	0.246	[-0.121, 0.844]	0.142
Low effort	-0.1617	0.296	[-0.741, 0.418]	0.584
Prior completion	3.6991	0.577	[2.569, 4.829]	0.000

low effort on the problem are highly likely to wheel-spin during the assignment where as there is a strong indication that students in the students exhibiting high effort are less likely to wheel-spin. The regression analysis is reported in Table 9.4.

We also examined the relationship between effort and assignment completion while controlling for prior completion. We found the model ($R^2 = 0.104$) found neither high nor low effort to be significant predictors of assignment completion although there was an indication that high effort is a predictor for assignment completion. Here, we found that the students who exhibit high effort will likely complete the assignment however the findings were not significant. The regression analysis is reported in Table 9.5.

9.8 Discussion and future works

Our analysis found that user behavior can be categorized into exhibiting low and high response times, which, in consideration of our exploratory analyses, we posit correspond to measures of high and low effort; we hypothesize from our findings that we are able to identify students applying effort as evidenced by the time taken and aspects of their subsequent action. With this definition of our metric, We found low effort

students to correlated strongly with wheel-spinning, even more so than the high effort students. This finding is a noteworthy contribution as it contradicts the intentional definition of wheel spinning behavior; many of the students exhibiting wheel spinning, in this way, appear to be spending little time and effort while working through their assigned work. We argue, and look to address in future work, that such students should not be considered as exhibiting wheel spinning and the definition of such behavior should be updated to consider these aspects of student work.

This work did not explore any interaction between effort and affect or other theories of behavior and engagement, but also may provide insights into student behavior across problems; the current analyses focuses at the sub-action level, and future works are planned to explore how our findings extend across an assignment. We are particularly interested in exploring the relationship between our measure of student effort and previously-developed measures of gaming behavior [263] while working on problems.

Other works have suggested that videos work better than hints in certain contexts [261], and future works intend to explore further if similar results may be better explained when accounting for student effort and attention devoted to the requested help. Additionally, in the future, we want to investigate if the effect in such studies is mediated by indicators of effort.

Similarly, the development of student models may benefit from further insights into student effort and engagement. Cognitive models such as that of Knowledge Tracing [80], for example, rely on correctness and incorrectness of student actions for modelling knowledge state, and we intuit that using a more continuous measure of effort might improve the performance of these types of cognitive models.

We implore researchers and developers to use our findings and exploration of effort to develop better measures and reports for teachers that consider effort in the assessment of students. We strive, in future works, to develop externally-validated measures of student engagement and effort toward these goals.

9.9 Conclusion

This paper presents evidence that provides new insights into user behavior pertaining to student help-seeking behavior. User response time can be categorized into users exhibiting high-effort and low-effort in their hint usage before taking the next action. We conducted exploratory analyses that helped to eliminate obvious systemic and performance confounds and still found distinguishable groups of students by the time devoted to hint requests. The response time decomposition work is an essential step in quantifying student effort while working on a problem as teachers often rely upon the amount of effort a student exhibits in conjunction with the student's problem-level correctness scores in gauging student progress while working

on their assignment.

We also explored the interaction between effort and wheel spinning as well as other student outcome measures. We found that lower effort students are highly correlated with wheel spinning behavior, contradicting the intended definition of the behavior; we argue that this is a significant finding as it attests to the fact that the definition of wheel-spinning needs further work as the current definition does not account for whether students are truly “spinning their wheels” by applying effort.

KNOWLEDGE TRACING OVER TIME: A LONGITUDINAL ANALYSIS

The use of Bayesian Knowledge Tracing (BKT) models in predicting student learning and mastery, especially in mathematics, is a well-established and proven approach in learning analytics. In this work, we report on our analysis examining the generalizability of BKT models across academic years attributed to "detector rot." We compare the generalizability of Knowledge Tracing (KT) models by comparing model performance in predicting student knowledge within the academic year and across academic years. Models were trained on data from two popular open-source curricula available through Open Educational Resources. We observed that the models generally were highly performant in predicting student learning within an academic year, whereas certain academic years were more generalizable than other academic years. We posit that the Knowledge Tracing models are relatively stable in terms of performance across academic years yet can still be susceptible to systemic changes and underlying learner behavior. As indicated by the evidence in this paper, we posit that learning platforms leveraging KT models need to be mindful of systemic changes or drastic changes in certain user demographics.

Proper citation for this chapter is as follows:

Lee, M.P., Croteau, E., Gurung, A., Botelho A.F., & Heffernan N.T. (2023). Knowledge Tracing Over Time: A Longitudinal Analysis. In The Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023).

10.1 Introduction

Modeling student knowledge and mastery of particular skills is a foundational problem to the domain of learning analytics and its intersections with education and artificial intelligence. The first proposed solution to the Knowledge Tracing (KT) problem, dubbed Bayesian Knowledge Tracing (BKT) by its creators [80], modeled knowledge as the mastery of multiple independent knowledge concepts (KCs, or skills) and estimated mastery through the use of a latent variable in a Hidden Markov Model. Student mastery of a skill is assumed to be a noisy representation of this latent variable, moderated by four parameters: a student's prior knowledge, the likelihood of mastering the skill through attempting a problem, the chance a student answers correctly by guessing, and the chance a student answers incorrectly by mistake. Future work augmenting

BKT attempted to improve model performance by modifying the assumptions of the initial model. For example, classical BKT models assume the acquisition of knowledge is unidirectional, from a state of non-mastery to a state of mastery. Relaxing this assumption and allowing for student knowledge to move bidirectionally between mastery and non-mastery resulted in models that more accurately predict student performance, and thus more accurately model student knowledge [294]. Further model extensions include allowing individual students to have personal prior knowledge rates [270] and giving individual questions their own guess and slip rates [271]. While other statistical models such as Performance Factors Analysis [275] showed initial promise, later advances in the domain of machine learning resulted in the creation of deep learning models to solve the problem of KT, utilizing a recurrent neural network in Deep Knowledge Tracing (DKT) [281] and self-attention in Self Attentive Knowledge Tracing (SAKT) [262]. However, BKT still serves as a useful way of modeling student knowledge due to the model's interpretability, especially in comparison to larger models [181]. BKT models require far fewer parameters to train in comparison to the deep-learning models even when BKT models incorporate the available extensions. If the performance of the model is a priority and the generalizability of the model is not guaranteed, then training new models in response to some population shift is advisable. Indeed, this is a common practice in online learning platforms when such shifts occur, such as the beginning of a new school year or the integration of a new curriculum. However, how do we know how often our KT models should be retrained?

More precisely, we wish to examine the performance of BKT models across time. Our analysis was guided by the following research questions:

- RQ1.** Do BKT models lose predictive power with time?
- RQ2.** Does the complexity of a KT model impact its generalizability through time?
- RQ3.** Do sudden shifts in student populations or behavior impact model performance?

To answer these questions, we gathered data collected through the ASSISTments platform across four school years from 2018-2022. We then compare model performance on data from the same year as training with model performance across years. Additionally, we posit that the COVID-19 pandemic caused a shift in student and teacher perception of technology for learning as there were no alternatives available to adopting technology in classrooms. As such we examine the shift in the learner behavior by examining the generalizability of KT models trained on pre-pandemic data to predict learning during the pandemic and vice versa. We begin by discussing the challenges to education posed by the COVID-19 pandemic, focusing on the rapid adoption of online learning tools during the pandemic. Next, we describe the data generation and

sampling process for our analysis. The student data available from ASSISTments across the four academic years establish a fair comparison of the KT models that is not susceptible to the size of the dataset since different academic years had varying number of users. We then describe the KT models used in our analysis and the approach we took in examining the generalizability of KT models. We compare model performance of classical BKT and BKT with forgetting models within the same academic year, across different academic years, and across the beginning of the pandemic, along with the impact of the forgetting parameter on model generalizability. We then discuss the implications of our findings on the implementation of KT models, and discuss the limitations of our analysis and their implications for future research.

10.1.1 COVID-19 Pandemic

The COVID-19 pandemic has presented many challenges to the delivery of education to students [124]. As many schools closed their doors, students were required to attend classes and complete coursework using online tools. This resulted in the rapid adoption of online learning platforms leading to a significant growth in the user base of platforms such as ASSISTments. This influx of new users likely introduces a more diverse group of students into school populations, since schools integrated various learning tools to support their students. Additionally, the sudden shift in the perception of technology and its use in teaching for many schools also present an interesting opportunity to explore the robustness and generalizability of KT models.

Given the wide-reaching changes to education caused by the COVID-19 pandemic, the impact these changes had on student learning requires more investigation. For the purposes of our analysis, we divided data gathered into two meta-groups: pre-pandemic and post-pandemic, with "post-pandemic" data merely denoting data that was gathered after the initial transition into online learning in mid-March 2020.

10.2 Related Work

Analysis of more complex inferential models used by MATHia found that models intended to detect "gaming the system" behaviors [23] trained on older data were significantly less precise on newer data [208]. It was found that more contemporary machine learning models designed to detect gaming experienced a greater performance decrease than classical, computationally simpler models. This phenomenon was called "detector rot" by its authors in reference to a similar phenomenon called "code rot" in which code performance decreases over time [175]. The analysis provided by [208] featured a comparison of models trained on data collected more than a decade apart, with models trained to solve a complex problem with a large feature space. We aim to contribute to the understanding of detector rot by examining model performance

Table 10.1: Dataset Information

Year	Total Rows	Total Assignments	Unique Students	% Correct
2018-2019	291,437	31,930	4,425	0.534
2019-2020	521,781	130,173	47,595	0.526
2020-2021	8,459,566	1,310,652	190,366	0.494
2021-2022	2,645,324	361,546	58,216	0.547

along more granular time steps, across dramatic population shifts, and with models solving a problem with a much smaller feature space.

10.3 Methods

10.3.1 Data Collection

Data for each school year was gathered from problem logs between the dates of September 1st and June 1st. Summer months were excluded as the student population during the summer can vary more drastically from year to year. The student cohort during some summers primarily consists of students requiring additional work to reach their credit requirements while other summers are filled with high achieving students working on extra credit. Problem level data from the typical academic year was then filtered based on several criteria in order to ensure different academic years were able to be directly compared. Comparison between two populations with little intersection in the skills being assessed would result in poor model generalizability based solely on underfitting. To ensure direct comparisons were possible and appropriate, we limited our underlying populations to problems sourced from the two most popular open-source math curricula available through OER [251] on the ASSISTments platform: EngageNY/Eureka Math and Kendall Hunt’s Illustrative Mathematics. From these two curricula, we calculated the top five hundred most commonly assigned problem sets across all four of our target years. The final populations we constructed before sampling were filtered by these top five hundred common problem sets, with the exception of the 2018-2019 school year. Data from this year was significantly more sparse than other years due to the introduction of a new implementation of the ASSISTments tutor, and as such we only applied the curriculum filter to this year. Since the introduction of the new tutor experience, student behavior has been logged in a consistent fashion.

Table 10.2: Feature List

Feature	Description
<i>user</i>	Unique student identifier
<i>assignment</i>	Unique identifier for an assignment
<i>correct</i>	0 if the student incorrectly applies skill, 1 otherwise
<i>start_time</i>	Timestamp of when the problem was started by the student
<i>problem</i>	Unique identifier for a problem
<i>curriculum</i>	Curriculum the problem originated from
<i>skill</i>	Skill being assessed by the current question
<i>attempt_number</i>	Counts which attempt on the problem this row represents

10.3.2 Student Modeling

Students in ASSISTments can make unlimited attempt when answering a problem until they answer it correctly, with the number of attempts a student takes to correctly answer a problem being recorded in problem-level data. The problem level data also includes information on the number of help requests and if the student requested for the answer to the problem. BKT attempts to predict student performance on attempts to apply a skill [80]. However, in the original problem level data, each student/problem interaction only has a single row. In an effort to encode information about how many attempts a student took to complete a problem, the original problem logs were used to create a dataset with each row representing a student's attempt to apply a skill. Additionally, if a student's final correct answer for a question came from a bottomed-out hint, explanation, or simply requesting the answer, the student's final correct answer was treated as an incorrect application of the skill. Information about the amount of data available for each year at the end of the filtering and encoding process can be found in Table 10.1, while a description of the available features present in all datasets can be found in Table 10.2. Ten samples of 25,000 assignment level data per year were generated for each year of the data. To investigate the effect of additional model parameters on model generalizability, two models were trained at each step: one with forgetting and one without. Other than this additional parameter, all training parameters were initialized in the same way. Models were constructed using pyBKT, a Python library for creating BKT models described by [21]. For analysis of within-year performance, a five-fold cross-validation was performed on each sample from the 10 samples, resulting in fifty measurements of AUC being taken for exploring model performance within the training year. For the inter-year performance analysis, the models were trained on one of the 10 random samples from a target

Table 10.3: BKT cross-year analysis

	18-19 Data	19-20 Data	20-21 Data	21-22 Data	Training Year Avg
18-19 Model		0.669	0.672	0.678	0.673
19-20 Model	0.682		0.729	0.714	0.709
20-21 Model	0.686	0.726		0.734	0.715
21-22 Model	0.690	0.724	0.748		0.721
Testing Year Avg	0.686	0.706	0.716	0.709	

Table 10.4: BKT+Forgets cross-year analysis

	18-19 Data	19-20 Data	20-21 Data	21-22 Data	Training Year Avg
18-19 Model		0.687	0.683	0.694	0.688
19-20 Model	0.686		0.740	0.730	0.719
20-21 Model	0.706	0.739		0.757	0.734
21-22 Model	0.708	0.736	0.766		0.735
Testing Year Avg	0.700	0.721	0.730	0.727	

year and evaluated on the other corresponding random samples from the other three years. This resulted in the generation of thirty measurements of AUC, since the model for each year was trained on 10 random samples and tested on 10 random samples from other three years resulting in 30 data points for the across year generalizability analysis. Finally, data from the 18-19, 19-20, and 20-21 years was split around the beginning of the COVID-19 pandemic (the precise date was March 12, 2020) and ten samples each containing 50,000 assignment level data were generated on each side of this split. The same process of five-fold cross-validation followed by a cross-year train/test analysis was performed on these pandemic samples.

10.4 Results

10.4.1 Robustness Over Time (RQ1)

Data gathered from our evaluations across academic years can be found in Tables 10.3 and 10.4, while the resulting means from our five-fold cross-validations plotted along with their 95% confidence intervals can be found in Figure 10.1. Rather unsurprisingly, the within year generalizability of the BKT models was high with the BKT + forgetting model always outperforming the classical BKT model. However the model generalizability when trained on one year and applied to other years varied across academic years: by

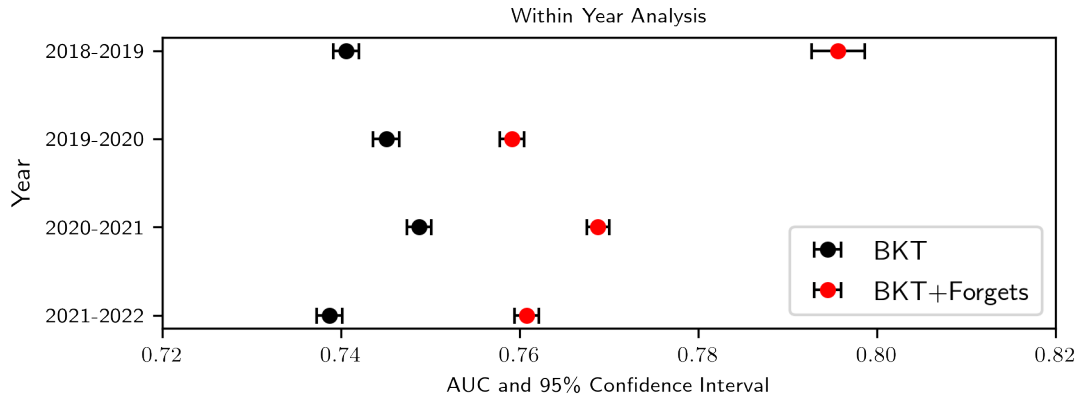


Figure 10.1: Means and 95% CIs for models trained and evaluated on the same year

comparing the training year averages provided in Tables 10.3 and 10.4, models trained on the 20-21 and 21-22 school years had higher average AUCs, while the 18-19 school year produced the least generalizable models. Similarly, different years were easier to generalize to than others, with the 18-19 school year having a much lower testing year average for both model types.

10.4.2 Complexity (RQ2)

One general observation seen from each of the analyses is that BKT+Forgets consistently outperforms classical BKT in terms of its predictive power as measured by mean AUC. Our findings strongly suggest the introduction of a forgetting parameter for each skill can be done with little chance of significantly harming a model's later generalizability.

10.4.3 Sudden Shifts: Pandemic Analysis (RQ3)

Data gathered from training and evaluating models before and after the COVID-19 pandemic can be found in Table 10.5, while these means and relevant confidence intervals were plotted in Figure 10.2. Models trained on data gathered before the pandemic had difficulties generalizing to post-pandemic data. Consider models evaluated on the post-pandemic dataset. The delta means between models trained on pre-pandemic data and post-pandemic data were 0.022 for classical BKT and 0.028 for BKT + forgets. This generalization problem also occurs when considering models evaluated on the pre-pandemic data, suggesting that KT models are susceptible to losses in predictive power following major shifts in underlying user populations.

As was true with the year-by-year data, the addition of a forgetting parameter to the classical BKT model significantly improves performance, even across the population shift. The use of model additions may im-

Table 10.5: Cross-Pandemic Analysis

Testing Period	Training Period	Model Type	Mean AUC	95%CE
Pre-pandemic	Pre-Pandemic	BKT	0.732	[0.731,0.733]
		BKT+Forgets	0.774	[0.772,0.776]
	Post-Pandemic	BKT	0.697	[0.696,0.698]
		BKT+Forgets	0.717	[0.715,0.720]
Post-pandemic	Pre-pandemic	BKT	0.727	[0.726,0.729]
		BKT+Forgets	0.742	[0.741,0.743]
	Post-pandemic	BKT	0.749	[0.748,0.750]
		BKT+Forgets	0.770	[0.769,0.771]

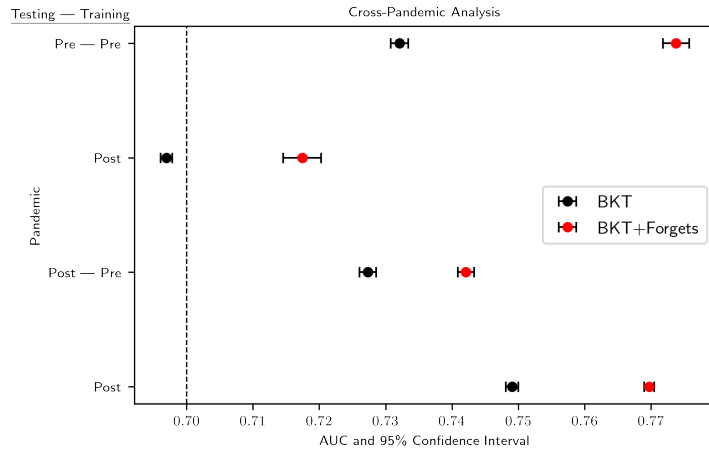


Figure 10.2: Means and 95% CIs for models trained on one side of the pandemic and trained on the other prove generalizability in a way that can withstand significant shifts in population and user behavior.

10.5 Discussion

In this paper, we explored the generalizability of KT models within and across academic years. The concept of "detector rot" [208] is a recent addition to how we understand inferential models and their applications in online tutoring platforms. With this analysis of how KT models perform over time, we intend to further explore the concept as it applies to KT models. Our exploration began by collecting data in a way that ensured the set of skills in each year's worth of data were comparable and then translating the raw problem level data into attempt-level representations of student performance. Models were evaluated both on the year in which they were trained (by a five-fold cross-validation), and on the other available years. We trained both classical

BKT models and models with a forgetting parameter to investigate how adding model parameters impacts model generalizability. We also divided our available data around the beginning of the COVID-19 pandemic to investigate the impact of sudden shifts in population size on model generalizability. We have a few key findings to report from these investigations. (a) In contrast to more sophisticated models, BKT's performance is relatively stable from year to year, indicating that the problem of detector rot is far less prevalent within the domain of KT. (b) The addition of forgetting parameters to BKT models consistently improves performance across multiple years of student population drift, and across more sudden changes of population. (c) Drastic changes in an online tutoring system's user base can impact BKT models' performance.

While our results indicate KT model stability over short-term population changes, our work is limited by several factors which future research could address. Our attempts to ensure each dataset contained a large overlap of skills could result in our models showing higher AUCs across time than comparable KT models would show in a product-scale system. Also, the 18-19 school year was particularly difficult for other models to generalize to. This is likely due to the sparsity of data for that year limiting our ability to filter by commonly assigned problem sets. Future work leveraging more data as ASSISTments continues to be used through time may give more insight as to why some years are easier for models to generalize to than others. Our analysis of RQ2 was also limited by only exploring how forgetting parameters impact generalizability. Future work incorporating more extensions to BKT, such as those described by [270] and [271], or utilizing more complex KT models like PFA [275] and DKT [281] is required to investigate trade-offs between model complexity and generalizability found in previous detector rot research [208]. Finally, while our analysis of RQ3 shows that BKT models had trouble generalizing across the beginning of the COVID-19 pandemic, the reasons for this could be numerous, including the sparsity of data pre-pandemic compared to post-pandemic or differences in student behavior after the pandemic began. Further analysis of how the COVID-19 pandemic impacted student behavior, possibly focusing on the transitional period from remote schooling back to in-person learning, could provide more insight into how student demographic changes affect KT models.

LIVE-CHART: LIVE INTERACTIVE VISUAL ENVIRONMENT FOR CREATING HEIGHTENED
AWARENESS AND RESPONSIVENESS FOR TEACHERS

Computer-Based Learning Platforms (CBLP) help enhance instruction and learning experiences through numerous dashboards, reports, and embedded tools and supports. Traditional CBLPs, however, generally do not facilitate real-time student-teacher interaction. We posit that real-time student-teacher interactions drive learning, especially amongst young students. In this paper, we meet this challenge through the design of a teaching augmentation tool that facilitates real-time interactions between teachers and students in physical and virtual learning environments. Following established methodologies in visualization design, we explore the goals of teaching augmentation tools broadly and propose and implement LIVE-CHART, a tool that exemplifies these factors. Using usability testing, we evaluate the utility of this system with teachers in real learning settings. Finally, we discuss the implications of LIVE-CHART's design and evaluation, including the need to identify modes through which educators may take action to support student learning and engagement in real-time as students are working on assigned content.

Proper citation for this chapter is as follows:

Gurung, A. (In preparation). LIVE-CHART: Live Interactive Visual Environment for Creating Heightened Awareness and Responsiveness for Teachers.

11.1 Introduction

Rapid progress in technology has provided us, as researchers and developers of educational technology, with the ability to optimize and innovate teaching-learning processes across classroom and virtual environments. The introduction of computer-based learning platforms (CBLP) into these settings has led to innovations to classrooms that have been disruptive to traditional practices in favor of more data-driven approaches that enhance the experience of both the teachers and students; research supports that these CBLPs have contributed to largely positive effects on student learning (c.f. [313]). In light of these impacts, it is important to consider the ways for which CBLPs are being used to drive such effects, as well as how they may better be leveraged to improve student learning and provide support for teachers.

A traditional classroom has two active agents: the students and the teachers. In most settings, it is the role

of the teacher to lead the learning experience in the classroom, while it is expected that students are largely expected to self-regulate their learning while working through assigned work; while supports are often made available, student learning follows a more self-paced paradigm in regard to their assigned work, especially in the context of homework. The motivation to optimize learning when students drive the learning process leads many CBLPs to primarily focus on improving the student experience. For example, many CBLPs facilitate automation through self-paced homework and classwork in the form of mastery-based assignments, mock tests/exams, immediate correctness feedback, as well as on-demand support in the form of hints, explanations, and scaffolding problems; all of these represent examples of student-facing functions provided by CBLPs.

Not all functions provided by many CBLPs are set to only benefit students, however, as there are many tools designed to support teachers as well. Leveraging the large amount of fine-grained data recorded as students interact with CBLPs, these systems are able to provide detailed reports on student performance that help teachers focus their attention to student needs. The automated grading and report generation has benefited teachers by saving time and providing instructional guidance. Within this, a branch of educational research has emerged to explore the use of computer-based visualization systems in summarizing student data to empower teachers in classrooms. These visualization techniques aim to address the needs of teachers by augmenting their capabilities with computational decision-making methods, helping them to optimize their time interpreting data and taking action; traditional classroom practices make this difficult as teachers spend time collecting, organizing, and interpreting data, detracting from their capacity to identify and address struggling students. These computer-based visualization systems are referenced as a subset of Classroom Orchestration [98, 312, 99, 361, 287, 168] or Teaching Augmentation (c.f. [9]) tools.

Teaching Augmentation (TA) refers to tools extending and complementing teachers' pedagogical abilities during ongoing classroom activities [9]. TA enables teachers to interface with their students in real-time through a virtual setting, providing insight into student progress and helping them to make data-driven actions that address learner needs at the individual, group, and class levels [168, 226, 8]. TA allows teachers to spend more time with students who need help [167, 225], enables teachers to distribute their time more fairly across students [8, 26], helps teachers identify students who are not engaging productively on the assignment [225, 17], and, through these, can help reduce learning achievement gaps across students [167]. Ultimately, TA facilitates teachers to effectively orchestrate their classes, amplifying their ability to recognize and address students' needs [17, 167, 225, 8, 26, 366, 85].

Globally, the COVID-19 pandemic caused a shift to remote and virtual learning paradigms, emphasizing many of the weaknesses in current learning tools and environments. Although CBLPs and technology, in

general, have aided schools in transitioning to virtual settings with varying degrees of success [14, 44, 330], the lack of TA tools is a notable deficiency of many CBLPs. Without TA tools, teachers may struggle to maintain strong interactions with students, and it may be similarly harder for teachers to gauge the effectiveness of their instruction without the feedback they would normally receive through interactions in face-to-face classroom environments.

However, while the strengths of TA tools, in general, are perhaps easy to identify, there are similarly limitations to current TA tools and supporting research that have led to difficulties of implementation and deployment in live contexts. For example, some proposed TA tools integrate with novel technology, including the use of motion and video sensors, in conjunction with virtual and augmented reality tools [167]; these tools are difficult to effectively scale due to monetary and ethical constraints (e.g. placing video cameras in classrooms), but similarly fail to extend to remote contexts. Alternatively, several TA tools do not support virtual scenarios. Prior works have documented teacher needs [168, 99, 312] and leveraged the principles of orchestration to inform their design [225]

Despite recent advancements in TA tools, there is an identified gap in existing research in abstracting and identifying the fundamental goals of such tools; the identification of these goals can help inform the development of future tools by helping to align innovation with the needs of teachers. The primary purpose of this paper is to explore the task abstractions for TA tools, in a general sense, and present a hierarchy of fundamental TA goals. We explore the goals of TA tools broadly, but then describe how these were considered in implementing LIVE-CHART, a teacher augmentation tool developed and tested with real teachers in varying educational settings. Additionally, we report the results of a usability study conducted to validate our understanding of the hierarchy of goals and how LIVE-CHART addresses teacher needs.

11.2 Theoretical Framework

While there are certainly variations, most traditional instructional paradigms are teacher-centric, where teachers drive the learning process. In these cases, educators with mastery over the subject matter project their knowledge through instruction, and students receive that knowledge through a combination of passive and active participation. Alternative strategies do exist such as flipped classrooms [2, 323, 41, 132], project-based learning [151, 36], peer learning [357, 51, 157], alternative discussion strategies [364], and innovative homework [137, 36]. However, these alternative strategies require students to be self-sufficient and proactive, which can be challenging for younger students [304, 303, 216, 243, 232]; we do not attempt to make a claim regarding the effectiveness of one instructional paradigm over another in this work, but we focus here

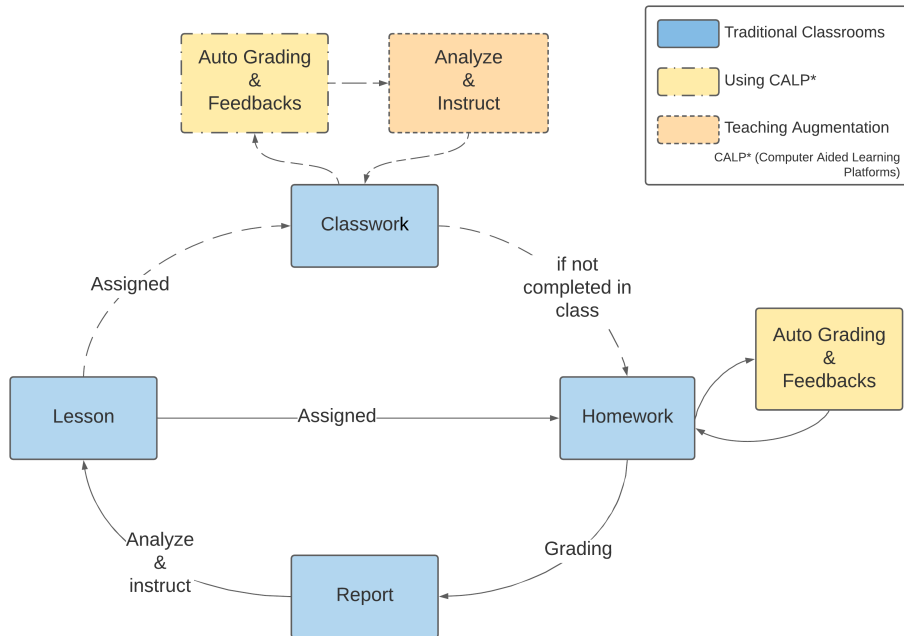


Figure 11.1: A visual representation of how the introduction of computers in the classroom has disrupted the learning experience in traditional classrooms.

on teacher-centric paradigms as these are arguably most common and also extend to other educational settings such as that of one-to-one and one-to-many tutoring. Upon deeper analysis, we posit one commonality within these methods: the student-teacher interaction drives learning. As such, we entrust the teacher to design their lesson plan to best suit the learning objectives of the subject matter. The design can vary across teachers depending upon their preferences and needs. Although each teacher is different and has a personalized lesson plan, we can generalize the structure of a class to a certain degree.

In figure 11.1, we present the cyclical structure of traditional classes (without the use of CBLPs) indicated by the four blue boxes of *Lesson*, *Classwork*, *Homework*, and *Report*; we propose that these four aspects of instruction form a simplified view of this cycle of instruction. The teacher first introduces the subject matter through a delivered lesson. The teacher can then assign classwork and/or homework to students to reinforce the lesson. After grading the assigned work, the teachers can analyze the resulting reports (i.e. student performance) to identify gaps in the students' knowledge and formulate a plan for addressing those gaps during the next lesson through remedial instruction. This traditional approach puts the teacher in control of the full process as they dictate the flow of the class by balancing learning needs, course requirements, and the timeline of content delivery. There are significant drawbacks to this approach as it is very demanding of teachers' limited time and resources as teachers grade the students' work, generate and analyze the report, identify student needs to make adjustments for the next lesson; this time limitation can be exacerbated with

the inclusion of remedial instruction that may detract from the time devoted to the next lesson. As this cycle can occur over multiple days, adjustments in scheduling to address identified gaps in students' understanding introduces the risk of a domino effect where the quality of the consecutive lessons might depreciate over time.

The introduction of CBLP, illustrated by the lighter yellow boxes denoted as *Auto Grading & Feedbacks* in figure 11.1, was a significant innovation in the teacher-centric instructional paradigm by decreasing the time needed for teachers to complete the cycle. CBLPs enhanced students' learning experiences via modalities such as formative feedback, automated and immediate grading of their answers, on-demand help in the form of hints, explanations, scaffolding problems, or worked-out examples, which allow for some supplemental instruction and remediation to be delivered as students work rather than a period of time afterward. On the other hand, CBLPs also help optimize teachers' time and resources by providing summative reports. The automated grading and report generation with additional data processing enabled the teachers to analyze and process the information to address students' needs much faster than without CBLPs. Arguably, these systems also support alternative peer-centric paradigms in a similar manner through, for example, an online forum model where students can ask questions and interact with each other [94, 90]. Research exploring the efficacy of CBLPs have shown evidence of a positive effect on learning in K-12 settings [247, 264, 313], and higher education [91, 225]. Still, the use of CBLPs alone still exhibits bottlenecks within this instructional cycle; teachers are unable to act on the information in real-time, requiring students to still complete their work before a teacher is able to utilize reports to identify content areas in need of further remediation and instruction. While these systems can help with this, current systems are limited in comparison to the experience and expertise of a teacher in regard to understanding the best actions to take in regard to addressing student needs.

The growing field of research, "Teaching Augmentation" [9] focuses on enhancing teachers' classroom abilities by leveraging the human-in-the-loop approach to aid teachers in accessing students' information in an actionable fashion. Teaching augmentation, depicted in orange as *Analyze & Instruct* at the top of Figure 11.1, is an approach to enhancing/augmenting teachers' ability to act in real-time within the classroom (or remotely). Using a TA tool can help teachers make data-driven decisions during class as students are working synchronously to address students' needs and gaps in their knowledge as they are working; this alleviates the bottleneck within the cycle by removing the delay between students working on assignments and the generation of reports. Implementation of effective visualization of student performance can help teachers gauge student progress and assess the overall performance of the class. Teaching augmentation has been shown to benefit students in classrooms [167, 8, 17, 226, 225], supporting their utility. Research and developments in the automated detection of various learning behaviors such as gaming the system [25, 24,

266, 265], attention [240, 100], and perseverance [35] can be leveraged within TA tools to provide deeper insights into student learning strategies, though there is still more research needed to understand the full extent to which such detectors can be used to inform action in live settings. Analyzing teacher needs and identifying the best way to augment their approach to instruction will help us identify the fundamental objectives of a TA tool, which can help guide the development of future systems to provide utility to teachers and students alike.

11.3 Examining Existing Teaching Augmentation tools

Various disciplines have explored teaching augmentation, including HCI [8, 26, 366, 182, 17], learning analytics [226, 365, 227], AI in education [167, 118, 360] and learning sciences [356]. We can make generalizations about various TA tools examining their implementation in classrooms, as dashboard interfaces [17, 225, 244, 94], as peripheral information displays [26, 8, 242], or as wearables [167, 37, 297]. A dashboard interface is invasive and requires active engagement from the teacher as the dashboard is ever-evolving to reflect student actions in real-time. While the dashboard design adds cognitive load on the teachers, it expedites their ability to identify students' needs and facilitate informed student-teacher interaction.

On the other hand, peripheral information displays are motivated by the calm technology [378] design principle that states that computer systems should passively engage with users by utilizing peripherals. While using peripherals is less cognitively demanding for teachers, they have lower bandwidth and only provide limited data. Wearables offer haptic feedback, which provides a distinct advantage over both dashboards and peripherals as it is cognitively less demanding while delivering more bandwidth. The benefits and drawbacks of peripherals vary depending on their type. For instance, wearables such as VR or AR goggles provide high bandwidth but are highly invasive, whereas smartwatches provide less bandwidth but are less invasive. Wearables also present feasibility challenges as they directly depend on the manufacturers' long-term support and updates; some wearables never cross the hurdle of concept-device and enter mass production, whereas others are too expensive to use in the everyday classroom. In light of the various modalities available to implement teaching augmentation, we analyzed four different projects exploring the TA tool in a real classroom.

Figure 11.2 shows the four tools that we further explored to gain insight into the various research projects that explored TA and analyze the benefits and drawbacks of their modalities. Figure 11.2A, Lumilo [167], uses mixed reality glasses as a TA tool. Figure 11.2B, Student Engagement Analytics Technology (SEAT) [17], uses tablets as a TA tool, and videos of individual students are analyzed to measure students' engagement. Figure 11.2C, Fireflies [8], uses peripheral devices in conjunction with a tablet as a form of TA. Figure 11.2D, MT Dashboard [225, 226], utilizes a tablet for TA within a class where students work on computers on table-

top mounted computers, and video of students and their performance on the classwork is analyzed to measure progress. In this paper, we utilized the dashboard design on a web application to develop our own teaching augmentation tool to remain consistent with the underlying infrastructure of a web-based CBLP.

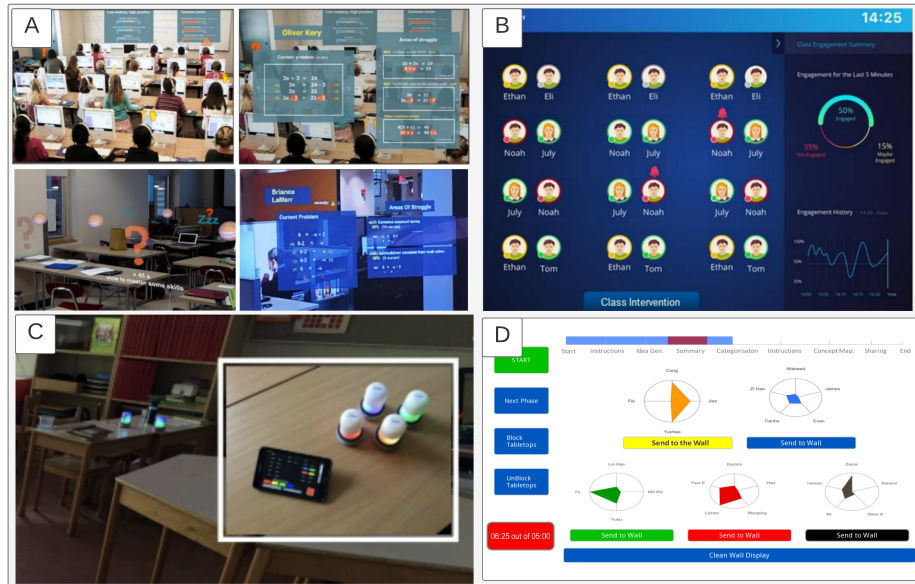


Figure 11.2: A visual representation of how the introduction of computers in the classroom has disrupted the learning experience in traditional classrooms. A. Lumilo, B. SEAT, C. Fireflies, and D. MTDashboard

A visual representation of how the introduction of computers in the classroom has disrupted the learning experience in traditional classrooms.

11.4 Task Abstraction

The development of goals and task abstraction for TA conducted for this work was an iterative process. In order to understand the requirements of a TA tool, we interacted with various domain experts, experienced teacher trainers, teachers, and researchers to understand the fundamental goals a TA tool needs to accomplish. Additionally we also analyzed the four TA tools presented in figure 11.2 as they provide valuable insights into the teacher needs and how to best address them. We divided the overall analysis into two parts. We began with goals analysis, where we developed a hierarchy of goals that a TA tool needs to facilitate. Next, we deconstructed the goals into subgoals. Subgoals give us high-level objectives that directly correlate with teacher needs. We leverage the concrete set of subgoals to enumerate visualization tasks in the TA tool that directly facilitate the teacher’s needs. We went through multiple iterations of goals and tasks analyses to further refine our findings and report on the final result.

Through LIVE-CHART, a TA tool that we have developed and present in this work, we have expanded

upon the affordances of Teaching Augmentation systems to facilitate an effective student-teacher interface as measured by the goals derived through our task abstraction procedure. LIVE-CHART provides teachers with insights into their students' performance on classwork, which influences the quality of student-teacher interactions as teachers address students' needs.

11.4.1 Goal Analysis

Table B.1 lists the goals and subgoals resulting from our analysis. The overarching goal of TA is to augment teachers' ability to interface with their students at an individual and class level. The goals help teachers identify the instances where they can address students one-on-one vs. as a group vs. the entire class. The first three goals, G1, G2, and G3, directly address teacher needs from various perspectives. The teachers need to analyze the entire class at a glance. Identify attendees, absentees, students who completed their classwork, and students who left their classwork incomplete. The teachers, at a glance, should be able to identify assignment progress for the entire class and individual students. When teachers are interested in comprehending the underlying causes behind student performance, TA should help teachers gain contextual insight. Infer the effort and attention a student put towards solving a problem and how close they came to solving it. Teachers have often expressed a wish to clone themselves so that they can help their students more effectively. TA helps teachers augment their abilities by helping them identify the students who require attention or students doing well. Goal 3 facilitates teachers to provide positive affirmations to students doing well and help students who require attention. Providing quantitative information on student performance is not enough as the teachers know their students and make nuanced inferences that a TA cannot make. For instance, student A taking their time working on the assignment might mean they need help, whereas student B taking the same amount of time might mean they are just thorough. These are nuances that come easily to a teacher but are challenging to quantify in TA.

11.4.2 Task Analysis

We identified the goals in Table B.1 related to abstract visualization tasks from Brehmer and Munzner's topology: the high-level family of Present tasks and the lower level family of Browse, Explore, and Identify tasks [49]. We provide the analysis of G1: Analyze assignment progress; the supplemental material contains the analyses for G2 and G3. G4 involves making inferences from the tasks in the other goals and hence does not have abstract visualization tasks of its own.

The high-level task that G1 supports is Consume: Present. The task can be further decomposed into auxil-

Table 11.1: Fundamental goals of a TA tool.

Generic Goals	
G1	Analyze assignment progress
a	Analyze assignment(classwork) progress for the entire class
b	Analyze assignment(classwork) progress for individual students
c	Identify problems where students struggled the most
G2	Identify underlying causes of differences in student performance
a	Use the student actions while working on the problems to gain contextual insight into student performance and effort
b	Infer the amount of effort and attention a student put in towards solving the problem
c	Identify problems where students struggled the most
G3	Enhance teacher performance
a	Identify students who are doing well or require attention
b	Help teachers facilitate equity in the classroom
G4	Discover nuances (Qualitative Inferences the teacher can make off of the quantitative data)

ary subtasks such as browse and identify depending on the tasks for the two subgoals. As shown in Table B.2, we can deconstruct G1, from Table B.1, into two subgoals, G1a Analyze assignment(classwork) progress for the entire class and G1b Analyze assignment(classwork) progress for individual students. Supporting G1a includes presenting the teachers with the students while they work on the assignment. Teachers can identify absentees, attendees, and the overall progress of attendees. The teachers can infer the students who are doing well on the assignment and moving along vs. the students who are slowing down and might need help from the teacher. For G1b, the teacher can infer the overall assignment progress of individual students and gauge the student performance in recent problems and the general status of the student’s presence in class.

11.5 Implementation through LIVE-CHART

In this section, we explore the data logged by a CBLP for students working through an assignment. We analyze the data and explore various approaches to leveraging the logged data to help achieve the fundamental goals of a TA from the previous section of task abstraction represented in table B.1. We explore this through the development and implementation of our own TA tool known as the *Live Interactive Visual Environment for Creating Heightened Awareness and Responsiveness for Teachers* (LIVE-CHART).

Table 11.2: Fundamental goals of a TA tool.

Tasks	
G1. a. Analyze assignment(classwork) progress for the entire class	
T1	Identify students who are absent and students who are present.
T2	Differentiate students who are working on the assignment, who have completed it, and who left without completing the assignment
T3	Analyze the difference between students who are quick vs. the students who take their time working on the assignment.
G1. b. Analyze assignment(classwork) progress for individual students	
T4	Analyze the problem correctness for the recent problems the student worked on
T5	Identify the percentage of the assignment the student completed
T6	Identify if the student is absent, working on the assignment, completed the assignment, or left without completing the assignment

11.5.1 Data Characterization

The underlying CBLP ¹ we are developing LIVE-CHART for tracks student actions while working on the assignment. The students can work on three different types of assignments: regular assignment, mastery-based assignment, and test/exam practice mode. The assignments consist of different types of problems to address different learning objectives. There are 11 different problem types: numeric, text, multiple-choice, ordering, and open response are some of the problem types. The system also provides students with feedback during the assignment. Students get feedback in the form of correctness feedback, common wrong-answer feedback, hints, explanations, and scaffolding problems. Correctness feedback notifies if the attempt was correct; common wrong-answer feedback helps students navigate misconceptions or silly mistakes that lead to the incorrect answer. Hints are designed to nudge the students towards the answer and have multiple parts, with the bottom out hint leading to an answer, whereas explanations are a singleton and give away the answer. The Scaffolding problems are subproblems that break the core concept of the main problem into subparts. The subparts guide the student to solve the problem step by step while helping them identify and address the gaps in their knowledge. While working on these problems, students can take different types of actions. Start a problem, attempt an answer, ask for help, complete the problem, and complete the assignment are examples of student actions. The CBLP logs 35 different student actions depending on the nature of the assignment

¹The name and some other characterizing details of this system are omitted for blinding purposes

and the types of actions a student takes while working on their assignment. The logging of information at the assignment and action levels forms a rich corpus of data per student working on the assignment. Teachers can find this information overwhelming to parse through and analyze in its current form as they need to process a large amount of data across several students. Through multiple design and development iterations, we analyzed logged data. The goal here was to identify the easily understandable aspects of data that can be actionable upon comprehension to the teachers. Our analysis segmented the entire data set into two major parts: assignment data and student data.

11.5.1.1 Assignment Data

Assignment data contains information on the assignment and its problems, the type of assignment, individual problem types, problem body, answer(s), common wrong answer, and the historical average score for the problem. The different types of problems require different response strategies when teachers address them during class. For instance, a common wrong answer in a multiple-choice problem must be treated differently from a common wrong answer for numeric problems. While multiple-choice requires the student to click on the option they think is correct, students need to do calculations before entering their answer for numeric problems. The average score on the problem provides teachers with insight into the problem's difficulty. As part of the design and engineering process, we constructed a data structure to expose the assignment data as JSON data streams for visualization. Table 11.3 provides a fragment of the JSON data stream in a tabular representation for three problems in an assignment. The first problem in table 11.3 is a Numeric problem the students need to solve. The answer for the problem is 12, and 18 is the common wrong answer. Historically the score for the problem is 0.33, indicating students have struggled on this problem. The next problem is an open response problem asking students to explain how they solved the first problem. The open response problems do not have an answer because students type in their reasoning. Teachers grade the open response problems, and there is no automated scoring for the open response. Table 11.3 does not include the assignment information: assignment type, start, and end date. Analysis of columns in table 11.3 provides insight into the data types: problem type is categorical, answer and common wrong answer are discrete and can be numeric or text depending on the problem type, problem body is discrete text associated with the problem and average score is continuous between the range [0.0, 1.0].

Table 11.3: A snippet of the problem information inside the data structure for assignment data.

Assignment Problems				
Problem Type	Problem Body	Answer	Common Wrong Answer	Average Score
Numeric	Solve for x, y=5: $5x + 3y = 75$	12	18	0.33
Open response	Solve for x, y=5: $5x + 3y = 75$ Explain your reasoning.	—	—	0.00
Multiple choice	What is the area of a rectangle?	$A = l * b$	$A = l * h$	0.66

11.5.1.2 Individual Student Data

Student data contains information on the students, student actions logged by the system, and problem logs representing the problem level information for the individual student. The problem level information for students represents the students' score per problem and the number of attempts the student made per problem. The system can log 35 action types when the student is working on a problem for action logs. We analyzed the data and filtered the 35 actions into a smaller set that is easily interpretable and actionable for the teachers. Problem start, attempt, help request, problem complete, and resume assignment represent some of the filtered actions. Each action also holds additional information for the action, such as the timestamp, action type, feedback id if applicable, correctness feedback for attempts. Table B has a tabular representation of a snippet of student actions working on a problem. The student started the problem with help available in the form of hints. After a while, the students made an attempt action as $A = l * b$. This action was their first attempt, and it was a correct response. The student got correctness feedback from the system, after which the student elected to move on to the next problem. Even after filtration, the data can be very overwhelming to teachers. Consider a teacher assigns a class of 10 students an assignment with 4 problems. Even if all the students answer the problems correctly, the teacher needs to analyze 160 filtered action logs to fully comprehend student performance to analyze the difference between quick vs. average pace students vs. students who take their time. This problem will only worsen if we account for the actions such as help requests, incorrect attempts, begin scaffolding. Analysis of columns in table 11.4 provides insight into the data types tracking student actions: timestamp is temporal, action type is categorical, the response to the problem is discrete and depends on the type of problem, help type is categorical, correctness is discrete(boolean), and the max attempt is ordinal.

From our data analysis, we concluded that action level data is at the nexus of interpretability and action-

Table 11.4: A snippet of the student data where the student answered the problem correctly after asking for a hint from the system.

Student Action Logs					
timestamp	Action Type	Help Type	response	correctness	Max Attempts
1561731328683	Problem Started	HINT	—	—	—
1561731365475	Attempt	—	$A = 1 * b$	true	1
1561731417168	Problem Completed	—	—	—	1
1561731418208	Continue To Next problem	—	—	—	—

ability. Action level information is temporal, and the actions dictate the flow of the information. The action level data is also structured to point to higher-order information at the problem and assignment level. Next, we describe LIVE-CHART’s visualization and interaction design, elaborating on how TA’s fundamental goals and data availability from the CBLP influenced our work.

11.6 Visualization and Interaction Design

The visualization of temporal data on a real-time dashboard was one of the most challenging aspects of this project. The traditional visualization of temporal and Spatio-temporal data relies on a combination of charts that represent various aspects of the dataset. Such visualizations, however, are designed for users who have a certain amount of domain expertise and are comfortable using the visualization. Unfortunately, our use-case does not allow us such affordance as we cannot ask teachers to dedicate their time to acquire expertise in a new domain as such visualization can introduce a steep learning curve that can deter adoption. As such, we revisited the previous implementations of TA in figure 11.4 and took inspiration from them.

11.6.1 Visualization Design

In order to address the fundamental goals of TA from table B.1, we divide the visualization into two major categories: a class view as shown in figure 11.3 and a student detail view as shown in figure 11.4.

11.6.1.1 Class View

Figure 11.3 shows the entire class with students arranged according to the classroom seating arrangement. At a glance, the teachers can identify absentees and attendees. From the attendees, teachers can identify students working on their assignments, students who completed their assignments, and students who left their assignments incomplete. Part (b) in figure 11.3 shows the classroom seating arrangement. Teachers

can also arrange their students alphabetically or in a per problem order based on the current problem of the student. The supplemental materials provide the alphabetical and per problem arrangement of the students. Part (d) in figure 11.3 shows the visualization of individual student progress at a high level as a card view. It helps teachers identify the student, their recent performance on the last 5 problem correctness, and the number of completed problems. We restrict the number of recent problems to 5 as we found the last 5 problems to represent recent student performance adequately. The number 5 is an arbitrary number that both the domain experts and teachers agreed was a fair compromise compared to showing student performance for all the problems. We also added visual encodings on the individual student views to represent absentees, incompletions, and completion. Additionally, we also added the visual encoding for requiring attention (red) and doing well (green) to the individual student views. We also integrated a “top bar”, figure 11.3(a), to help surmise students doing well and requiring attention to help teachers prioritize those students. We did not explore any complex models to classify students into the two bins of doing well and requiring attention. Currently, LIVE-CHART classifies students who answer three consecutive problems correctly as doing well and students who answer three problems incorrectly as requiring attention. Similar to the last 5 problems for recent performance, the domain experts and teachers agreed that 3 consecutive correctness and incorrectness was a fair compromise in classifying student progress.

11.6.1.2 Student Detail View

When a teacher identifies a student for an in-depth analysis to gain some contextual insight into their progress, the teacher can click on the student card view to view the “student detailed view”. Figure 11.4 shows a detailed view for “11 Student”; part (a) indicates the student requires attention, part (b) shows the problems the student has completed, and part (d) shows a detailed breakdown of student actions. The figure shows the action logs for the last two problems the student attempted. The student made several incorrect attempts before finally answering the second last problem, whereas their first action was to ask for the answer for the last problem. The change in student behavior provides the teacher with valuable insight that can inform their instruction. Combining teachers’ knowledge of the student and the contextual insight from LIVE-CHART can lead to more productive student-teacher interactions. The teacher can acknowledge the student’s effort on the second last problem and provide instruction and encouragement that can positively impact the student. The positive impact can manifest in various forms: a sense of belonging from how invested their teacher is in their work, a sense of purpose for acknowledgment of their effort, and the hands-on instruction from the teacher might address the gaps in the student knowledge. Conversely, the teacher’s

knowledge of the student and the contextual insight might help the teacher identify the sequence of actions as gaming behavior. The teacher can intervene and instruct the student to stop the behavior and work on the material. The teacher intervention can lead to a different positive impact as it encourages accountability amongst students as teachers discourage such behaviors.

The two visualizations: class view and student detail view, encapsulates the fundamental goals of TA presented in table B.1. Figure 11.3 (b) and (d) addresses goal G1 by enabling teachers to analyze the assignment progress of the entire class as well as specific students. Figure 11.4 (b) and (d) address goal G2 by facilitating teacher to gain contextual insights into the student performance on the problems. Finally, figure 11.3 (a) and figure 11.4 (a) address goal G3 of highlighting students who are doing well or require attention to provide positive affirmations and help, respectively. Combining teachers' knowledge of their students and the visualization can help teachers discover nuances and make inferences they cannot make through visualization alone, addressing goal G4.

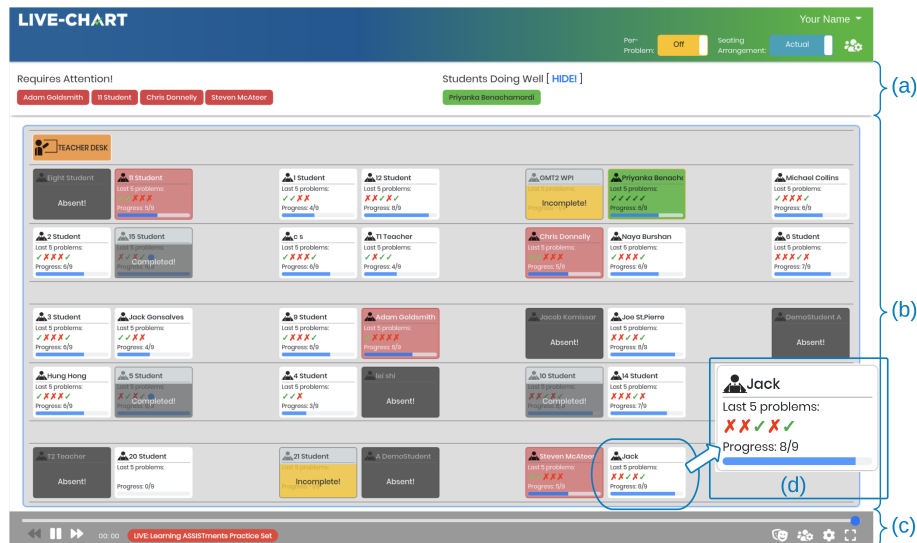


Figure 11.3: Class view: visualization of students as they work on their classwork in real-time. (a)priority dashboard that highlights students doing well or requiring attention in class, (b) students visualized according to their classroom seating arrangement, (c) The controls: teachers can manipulate the visualization timeline as the data is temporal, and (d) individual student view representing high level information for teacher to infer progress.

Class view: visualization of students as they work on their classwork in real-time. (a)priority dashboard that highlights students doing well or requiring attention in class, (b) students visualized according to their classroom seating arrangement, (c) The controls: teachers can manipulate the visualization timeline as the data is temporal, and (d) individual student view representing high level information for teacher to infer current status of student.

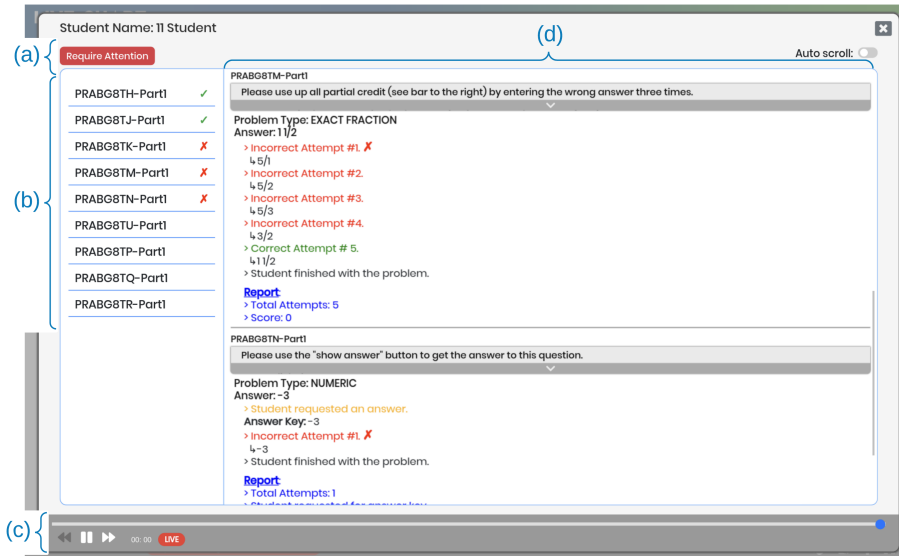


Figure 11.4: Student Detail View: visualization of individual student’s work with detailed per-problem information in real-time. (a) dashboard indicating if the student currently has been flagged for requiring attention or doing well, (b) problem level information representing student performance in prior problems, (c) The controls: teachers can manipulate the visualization timeline as the data is temporal (d) detailed breakdown of the problem and the actions a student took while working on the problem.

Student Detail View: visualization of individual student’s work with detailed per-problem information in real-time. (a) dashboard indicating if the student currently has been flagged for requiring attention or doing well, (b) problem level information representing student performance in prior problems, (c) The controls: teachers can manipulate the visualization timeline as the data is temporal (d) detailed breakdown of the problem and the actions a student took while working on the problem.

11.6.2 Interaction Design

Designing an interface to manipulate the temporal visualization while ensuring that the UI was intuitive with minimal learning requirements was the most challenging aspect of LIVE-CHART. The desire to design an intuitive UI that leads to a seamless UX led us to explore radial dials, timelines with sliders, start and current time with text inputs where a teacher can directly enter the time and jump to a previous instance. While some of these ideas were difficult to implement, others were downright clunky and difficult to justify. After several iterations of design processes, we settled on replicating the UI/UX of a video player. The inspiration to replicate the video player came from observing the UX of “live streaming videos.” Most users are familiar with the controls of a video player, giving us the convenience of implementing a controller that allows teachers to go back in time and periodically analyze their students’ past performance. Figure 11.3 (c) and figure B.1 (c) shows the final implementation of the interaction module. There is a unidirectional temporal

link between the class view and student detail view where if the teacher goes back in time and analyzes the class view, then the student detail view also opens up at the exact time the teacher selected. The converse, however, is not valid as teachers might go back in time to analyze past actions of a student in the “student detailed view” without intending to go back in time for the entire class. As part of the interaction design, we also allow teachers to implement different types of classroom seating arrangements that can help teachers analyze the effectiveness of different seating arrangements and identify groups of students who perform well during assignments when seated next to each other vs. students who struggle with assignments when placed next to each other. The interaction design to help teachers create their own seating arrangement is shown in figure B.1.

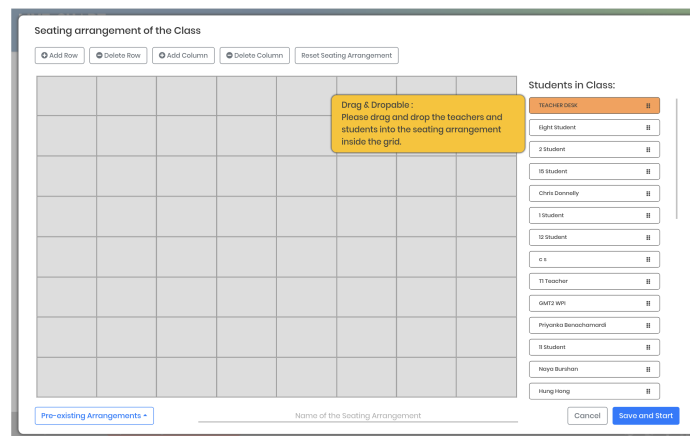


Figure 11.5: Seating Arrangement: teachers can arrange the students in the class to reflect the seating arrangement of the class.

Seating Arrangement: teachers can arrange the students in the class to reflect the seating arrangement of the class.

11.7 Usability Study

The overarching goal of LIVE-CHART is to encourage effective student-teacher interactions. We primarily designed LIVE-CHART to support teachers during an in-person class; however, we also extended the design to include alphabetical arrangement and per-problem views to facilitate teachers during virtual classes or tutoring sessions. Our original plan of a usability study for in-person classes was ultimately canceled due to circumstances pertaining to the COVID-19 pandemic, so we conducted our usability study in a virtual class setting; this is important to mention as there are potentially limitations that were not revealed through this paradigm, all of which are intended to be explored and addressed through future work. The usability study used a combination of surveys and semi-structured interviews. For the usability study, we recruited 8

teachers from 6 different states across the United States. The teachers in the study were not part of the design and development process. They taught mathematics to grade 7 and 8 students. Figure 11.6 provides insight into the average class sizes of our recruited teachers. We group the class sizes into three groups: small(≤ 24), regular[22, 28], and large(≥ 28). The teachers reported that they use the underlying CBLPs at least once a week for homework(7), classwork(1), standardized test preparation(1), and assessment(1).

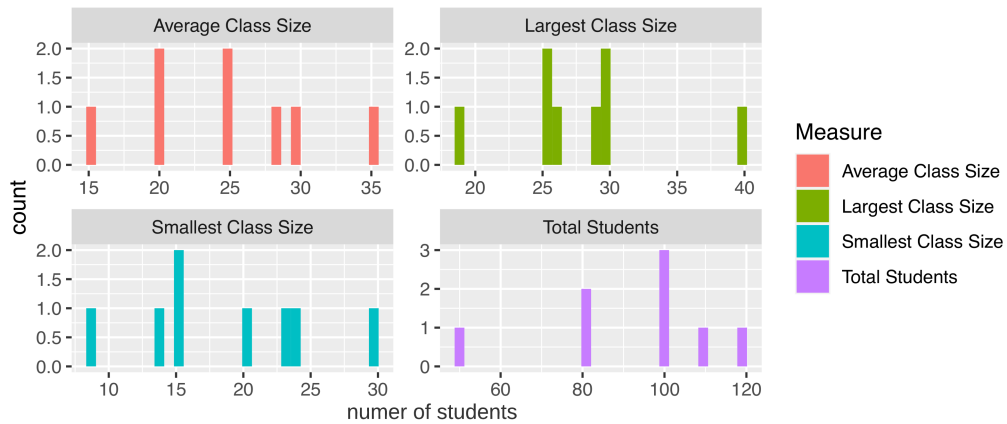


Figure 11.6: The class sizes of the teachers in the study.

The class sizes of the teachers in the study

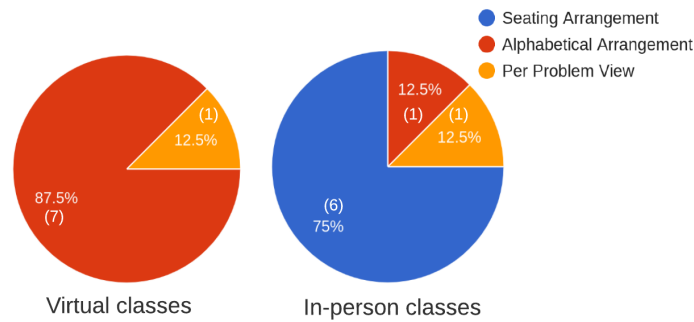


Figure 11.7: Teacher preference for student arrangement during virtual and in-person classes.

Teacher preference for student arrangement during virtual and in-person classes

We conducted 3 virtual workshops with surveys at the end to familiarize the teachers with LIVE-CHART. The first workshop was a presentation on the idea of LIVE-CHART to collect teacher feedback. Of the 8 teachers, 6 were open to using LIVE-CHART immediately, whereas 2 wanted to explore it further before trying it in their classrooms.

The second workshop was hands-on, and we asked the teachers to mimic student behaviors and work on

an assignment while we demonstrated LIVE-CHART by tracking their progress. Even though the teachers had only witnessed us use LIVE-CHART, everyone gave positive feedback. One teacher is quoted as saying *“I think it is very user-friendly, particularly in that there are different options teachers can choose from to determine what works best for them. I think for me when in my classroom, the seating arrangement function is phenomenal to see visually how each table of students did on a particular assignment or question. Virtually, after thinking about it, I actually think the alphabetical arrangement would be most useful to me. Live in the classroom I like the “horse race” option the best. The flexibility adds to user-friendliness. The visual aspects (checks and x’s) and color coding are very easy to understand at a quick glance.”* The “horse race” option refers to the per-problem view; it was an internal term one of our stakeholders had given the per-problem view and the term caught on with the teachers participating in the workshop.

The third session had a role reversal. We behaved as students, and the teachers monitored our progress using LIVE-CHART. After the third workshop, we collected teacher feedback on the usability of the seating arrangements represented in figure 11.7. For virtual classes, 7 teachers thought alphabetical arrangement worked best for them, and 1 teacher thought per-problem view worked best for them. For in-person classes, 6 teachers thought seating arrangement was the best fit, 1 teacher thought alphabetical arrangement was the best fit, and 1 teacher thought the per-problem view was the best fit.

11.7.1 Real Time usage

After the workshops, the teachers used LIVE-CHART once a week to monitor student progress during classwork for three consecutive weeks. As the teachers taught more than one class, we randomized at a class level to compare and contrast the usability of LIVE-CHART when the students were working on their classwork. One school closed for a week in the middle of the study, citing a lack of resources and support for the teachers. A different school had scheduled vacation during the third week of the study. For both situations, we coordinated with the teachers and continued the study when school resumed. After completing the third week, we met the teachers and conducted individual semi-structured interviews to get teacher feedback on LIVE-CHART usability. Due to these complications and the otherwise limited scale of this study, we report the results here in a more qualitative manner as the purpose was to assess the usability of the tool as opposed to its efficacy. From the feedback gained from this study, we are actively developing improvements to address highlighted limitations while offering the described strengths to help inform others pursuing similar development efforts. A more formal efficacy study is planned for future work.

11.7.2 Takeaways

There were several notable takeaways from the usability study that may help inform the development of TA tools in a broad sense. These are described here as they pertained to LIVE-CHART, but are presented along with broader suggestions for approaching implementations in other platforms.

The most notable limitation expressed by teachers was the timing of information. As students worked through assignments, LIVE-CHART exhibits a 20-second delay between the student taking an action and that being shown on the interface. There are several systemic reasons for this delay (some of which can be easily adjusted), but this was an unexpected limitation. Some teachers struggled with the delay, as it resulted in some unforeseen scenarios: for instance, say a student answered three problems incorrectly, and LIVE-CHART flags the student for requiring attention. When the teacher notices a student requiring attention and goes to help them on problem N , the student might be working on problem $N+1$ or $N+2$ if they solved $N+1$ within 20 seconds. As without the TA tool teachers would normally need to wait until students finished the assignment to assess their progress, it was surprising that the delay raised these issues. This does suggest that teachers 1) want to address students immediately, and 2) are seemingly able to adjust their normal instructional paradigm to include real-time action. Further studies are needed to better understand the threshold of what delay is tolerable, particularly as incorporating more advanced detectors of student performance and behavior are likely to require processing time within such a tool as well.

Some teachers reported starting with the per-problem view initially, however by the third week, everyone reported using the alphabetical arrangement during the class. We did not give the teachers any instruction about the time they needed to allocate for classwork to their students. All the teachers reported that they used LIVE-CHART in the last 15 minutes of the class after completing the lesson. In this case, we were similarly surprised that most teachers settled on using the alphabetical arrangement, though we suspect that the virtual setting of the study contributed largely to this; it is anticipated that other views may provide other utilities in a physical setting.

LIVE-CHART also has a playback feature where teachers can replay their classes to analyze student performance. Although the playback mode was not part of the study, 7 of the 8 teachers reported using playback mode more than once a week to review assignments. The teachers also reported how the playback mode helped make data-driven decisions as they could take more time analyzing student actions during playback as there were no time constraints. In this way, the playback feature was used in a similar manner as other reports in non-TA CBLPs, where student performance is viewed in a summative manner after students have completed the work. Teachers seemingly found the temporal aspects of the tool to provide insights that were

not easily visualized through more static reports. This finding has led us to begin exploring other ways of merging summative information into the playback tools in an attempt to further increase utility (and perhaps negating the need for teachers to view two types of report to view the information they desire).

The majority of the teachers reported on LIVE-CHART improving student accountability and engagement once the students realized their teacher could monitor their progress. This finding aligned with previous work which found that, even without enhanced analytics, student perceptions of being potentially monitored led to increased engagement [167]. A teacher also reported how they leveraged LIVE-CHART to identify the quicker students who completed their classwork and were assigned additional work for extra credit.

11.8 Future work and Open Questions

While we have indicated several directions for future work in previous sections, there are several other aspects that can be addressed with further research. Within this, there are many open questions that our study, and those in prior works, have left unanswered. These are introduced and described in this section in the hopes that these open questions may help guide future studies for LIVE-CHART as well as motivate other researchers and developers to pursue these in other contexts as well.

Prior implementations of TA have explored the approach of leveraging higher-order information on student behavior, student affect [167], and engagement [17, 225] as opposed to leveraging problem-level information as was done in LIVE-CHART. It is not evident which approach works the best: higher-order information on student behavior, granular problem level information, or a combination of the two. This unknown further raises the additional question of whether the implementation of student behavioral detectors influence teachers' perceptions of LIVE-CHART and TA in general? Will it develop reliance and trust or perhaps create mistrust and skepticism?

One aspect of the student-teacher interaction that we did not explore in LIVE-CHART was student autonomy. Past research has explored the allowance of student autonomy in the classroom through TA [8]. Does the autonomy facilitate persistence [236, 316, 209, 171] and grit [105, 164], resulting in better learning, or does it lead to abuse of the system by students electing to avoid interacting with their teachers?

An aspect of classwork that the current design of LIVE-CHART does not facilitate is the ability to collaborate with peers as interactions in the classroom do not simply occur between students and teachers; it also occurs amongst students. Collaboration and working in groups can also positively affect intrinsic factors, such as a sense of belonging, drive to keep up with peers, and a sense of achievement.

One of our takeaways about the integration of LIVE-CHART into the classroom was that teachers used

it during the last 15 minutes of the class. There is an open-ended question as to the effectiveness of LIVE-CHART and time. Is it better to have a lesson plan with a small chunk of time allocated to classwork? Or is it better to teach a topic across consecutive days and separate an entire period once a week as a classwork period where students work on the problem from the concepts learned during the week?

11.9 Conclusion

In this work, we examine, decompose, and explore the fundamental goals and tasks that a TA tool needs to integrate into a classroom successfully. We then used this goal set to develop and implement LIVE-CHART, serving as a proof-of-concept and testbed for these ideas. We evaluate this approach by means of a usability study on LIVE-CHART with teachers in real learning environments. We report qualitative feedback from our teachers collected through semi-structured interviews and surveys and offer several guiding directions to support the development of TA tools broadly. This work also aims to provide researchers and developers with insights into a TA tool's design, acting in-part as a "lessons learned" to support teachers and students through the development of these and similar tools.

REFERENCES

- [1] Abdi, H., “Holm’s sequential bonferroni procedure”, *Encyclopedia of research design* **1**, 8, 1–8 (2010).
- [2] Abeyssekera, L. H. D. J. and P. Dawson, “Motivation and cognitive load in the flipped classroom: definition, rationale and a call for research”, *Higher Education Research & Development* **34**, 1, 1–14 (2015).
- [3] Akay, H., D. Soybaş and Z. Argün, “Problem kurma deneyimleri ve matematik öğretiminde açık-uçlu soruların kullanımı”, *Gazi Üniversitesi Kastamonu Eğitim Dergisi* **14**, 1, 129–146 (2006).
- [4] Aleven, V., B. McLaren, I. Roll and K. Koedinger, “Toward tutoring help seeking”, in “International Conference on Intelligent Tutoring Systems”, pp. 227–239 (Springer, 2004).
- [5] Aleven, V., E. Stahl, S. Schworm, F. Fischer and R. Wallace, “Help seeking and help design in interactive learning environments”, *Review of educational research* **73**, 3, 277–320 (2003).
- [6] Allcott, H. and J. B. Kessler, “The welfare effects of nudges: A case study of energy use social comparisons”, *American Economic Journal: Applied Economics* **11**, 1, 236–76 (2019).
- [7] Allen, L. K., M. E. Jacovina and D. S. McNamara, “Computer-based writing instruction.”, Grantee Submission (2016).
- [8] An, P., S. Bakker, S. Ordanovski, R. Taconis, C. L. Paffen and B. Eggen, “Unobtrusively enhancing reflection-in-action of teachers through spatially distributed ambient information”, in “Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems”, pp. 1–14 (2019).
- [9] An, P., K. Holstein, B. d’Anjou, B. Eggen and S. Bakker, “The ta framework: Designing real-time teaching augmentation for k-12 classrooms”, in “Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems”, pp. 1–17 (2020).
- [10] Anderson, J. R., A. T. Corbett, K. R. Koedinger and R. Pelletier, “Cognitive tutors: Lessons learned”, *The journal of the learning sciences* **4**, 2, 167–207 (1995).
- [11] Anderson, S. A., “Synthesis of research on mastery learning.”, URL <https://files.eric.ed.gov/fulltext/ED382567.pdf> (1994).
- [12] Anthology, “Blackboard learning”, URL <https://www.blackboard.com/en-eu/teaching-learning/learning-management> (1997).
- [13] Antinyan, A. and Z. Asatryan, “Nudging for tax compliance: A meta-analysis”, *ZEW-Centre for European Economic Research Discussion Paper* , 19-055 (2019).
- [14] Armstrong-Mensah, E., K. Ramsey-White, B. Yankey and S. Self-Brown, “Covid-19 and distance learning: Effects on georgia state university school of public health students”, *Frontiers in Public Health* **8**, 576227–576227 (2020).
- [15] Arroyo, I., J. E. Beck, C. R. Beal, R. Wing and B. P. Woolf, “Analyzing students’ response to help provision in an elementary mathematics intelligent tutoring system”, in “Papers of the AIED-2001 workshop on help provision and help seeking in interactive learning environments”, pp. 34–46 (Cite-seer, 2001).
- [16] Arroyo, I., B. P. Woolf, W. Burelson, K. Muldner, D. Rai and M. Tai, “A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect”, *International Journal of Artificial Intelligence in Education* **24**, 4, 387–426 (2014).
- [17] Aslan, S., N. Alyuz, C. Tanriover, S. E. Mete, E. Okur, S. K. D’Mello and A. Arslan Esme, “Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms”, in “Proceedings of the 2019 chi conference on human factors in computing systems”, pp. 1–12 (2019).

- [18] Association, N. G. *et al.*, “Common core state standards”, Washington, DC (2010).
- [19] Babad, E. Y., “Expectancy bias in scoring as a function of ability and ethnic labels”, *Psychological Reports* **46**, 2, 625–626 (1980).
- [20] Babad, E. Y., M. Mann and M. Mar-Hayim, “Bias in scoring the wise subtests.”, *Journal of Consulting and Clinical Psychology* **43**, 2, 268 (1975).
- [21] Badrinath, A., F. Wang and Z. Pardos, “pybkt: an accessible python library of bayesian knowledge tracing models”, arXiv preprint arXiv:2105.00385 (2021).
- [22] Baker, R. S., A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell and S. Giguere, “Contextual slip and prediction of student performance after use of an intelligent tutor”, in “International conference on user modeling, adaptation, and personalization”, pp. 52–63 (Springer, 2010).
- [23] Baker, R. S., A. T. Corbett, K. R. Koedinger and A. Z. Wagner, “Off-task behavior in the cognitive tutor classroom: When students’ game the system””, in “Proceedings of the SIGCHI conference on Human factors in computing systems”, pp. 383–390 (2004).
- [24] Baker, R. S., A. T. Corbett, I. Roll and K. R. Koedinger, “Developing a generalizable detector of when students game the system”, *User Modeling and User-Adapted Interaction* **18**, 3, 287–314 (2008).
- [25] Baker, R. S., A. T. Corbett and A. Z. Wagner, “Human classification of low-fidelity replays of student actions”, in “Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems”, vol. 2002, pp. 29–36 (2006).
- [26] Bakker, S., E. van den Hoven and B. Eggen, “Fireflies: physical peripheral interaction design for the everyday routine of primary school teachers”, in “Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction”, pp. 57–64 (2013).
- [27] Bangert-Drowns, R. L., C.-L. C. Kulik, J. A. Kulik and M. Morgan, “The instructional effect of feedback in test-like events”, *Review of educational research* **61**, 2, 213–238 (1991).
- [28] Bangert-Drowns, R. L., J. A. Kulik and C.-L. C. Kulik, “Effects of frequent classroom testing”, *The journal of educational research* **85**, 2, 89–99 (1991).
- [29] Bar, T., V. Kadiyali and A. Zussman, “Grade information and grade inflation: The cornell experiment”, *Journal of Economic Perspectives* **23**, 3, 93–108 (2009).
- [30] Baral, S., A. F. Botelho, J. A. Erickson, P. Benachamardi and N. T. Heffernan, “Improving automated scoring of student open responses in mathematics.”, *International Educational Data Mining Society* (2021).
- [31] Baral, S., K. Seetharaman, A. F. Botelho, A. Wang, G. Heineman and N. T. Heffernan, “Enhancing auto-scoring of student open responses in the presence of mathematical terms and expressions”, in “International Conference on Artificial Intelligence in Education”, pp. 685–690 (Springer, 2022).
- [32] Bastin, C. and M. Van der Linden, “The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging.”, *Neuropsychology* **17**, 1, 14–24 (2003).
- [33] Bates, D., M. Mächler, B. Bolker and S. Walker, “Fitting linear mixed-effects models using lme4”, *Journal of Statistical Software* **67**, 1, 1–48 (2015).
- [34] Beck, J. E., K.-m. Chang, J. Mostow and A. Corbett, “Does help help? introducing the bayesian evaluation and assessment methodology”, in “International Conference on Intelligent Tutoring Systems”, pp. 383–394 (Springer, 2008).
- [35] Beck, J. E. and Y. Gong, “Wheel-spinning: Students who fail to master a skill”, in “International conference on artificial intelligence in education”, pp. 431–440 (Springer, 2013).
- [36] Bennett, R. E., “Cbal: Results from piloting innovative k–12 assessments”, *ETS Research Report Series* **2011**, 1 (2011).

- [37] Berque, D. A. and J. T. Newman, “Glassclass: Exploring the design, implementation, and acceptance of google glass in the classroom”, in “International Conference on Virtual, Augmented and Mixed Reality”, pp. 243–250 (2015).
- [38] Bhatnagar, S., N. Lasry, M. Desmarais and E. Charles, *DALITE: Asynchronous Peer Instruction for MOOCs*, vol. 9891 of *Lecture Notes in Computer Science*, p. 505–508 (Springer International Publishing, Cham, 2016), URL http://link.springer.com/10.1007/978-3-319-45153-4_50.
- [39] Bhatnagar, S., N. Lasry, M. Desmarais and E. Charles, “Dalite: Asynchronous peer instruction for moocs”, in “Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13-16, 2016, Proceedings 11”, pp. 505–508 (Springer, 2016).
- [40] Biancarosa, G. and C. E. Snow, *Reading next: A vision for action and research in middle and high school literacy: A report from Carnegie Corporation of New York* (Alliance for Excellent Education, 2004).
- [41] Bishop, J. L., J. Bishop, M. A. Verleger, E.-R. Aeronautical and D. Beach, “The flipped classroom: A survey of the research”, ASEE Annual Conference and Exposition, Conference Proceedings (2013).
- [42] Bjork, R. A., “Memory and metamemory considerations in the training of human beings”, *Metacognition: Knowing about knowing* **185**, 7.2, 185–205 (1994).
- [43] Bloom, B., “The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring”, *Review of Educational Research* **13**, 6, 4–16 (1984).
- [44] Bookser, B. A., M. Ruiz, A. Olu-Odumosu, M. Kim, S. N. Jarvis and J. A. Okonofua, “Context matters for preschool discipline: Effects of distance learning and pandemic fears.”, *School psychology* (Washington, D.C.) (2021).
- [45] Bosch, N., S. K. D’mello, J. Ocumpaugh, R. S. Baker and V. Shute, “Using video to automatically detect learner affect in computer-enabled classrooms”, *ACM Transactions on Interactive Intelligent Systems (TiiS)* **6**, 2, 1–26 (2016).
- [46] Botelho, A. F., R. S. Baker, J. Ocumpaugh and N. T. Heffernan, “Studying affect dynamics and chronometry using sensor-free detectors.”, *International Educational Data Mining Society* (2018).
- [47] Botelho, A. F., S. Baral, J. A. Erickson, P. Benachamardi and N. T. Heffernan, “Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics”, *Journal of Computer Assisted Learning* (2023).
- [48] Brehmer, M. and T. Munzner, “A multi-level typology of abstract visualization tasks”, *IEEE transactions on visualization and computer graphics* **19**, 12, 2376–2385 (2013).
- [49] Brehmer, M. and T. Munzner, “A multi-level typology of abstract visualization tasks”, *IEEE Transactions on Visualization and Computer Graphics* **19**, 12, 2376–2385 (2013).
- [50] Brennan, D. J., “University student anonymity in the summative assessment of written work”, *Higher Education Research & Development* **27**, 1, 43–54 (2008).
- [51] Broadbent, J. and W. L. Poon, “Self-regulated learning strategies & academic achievement in online higher education learning environments: a systematic review”, *Internet and Higher Education* **27**, 1, 1–13 (2015).
- [52] Brown, H. D. and P. Abeywickrama, *Language Assessment: Principles and Classroom Practices*. (Pearson Education, New York, USA, 2004).
- [53] Brown, J. S. and R. R. Burton, “Diagnostic models for procedural bugs in basic mathematical skills”, *Cognitive science* **2**, 2, 155–192 (1978).

- [54] Brown, J. S. and K. VanLehn, “Repair theory: A generative theory of bugs in procedural skills”, *Cognitive science* **4**, 4, 379–426 (1980).
- [55] Bruning, R. H., G. J. Schraw and M. M. Norby, *Cognitive psychology and instruction* (Pearson, 2011).
- [56] Burnette, J. L., M. V. Russell, C. L. Hoyt, K. Orvidas and L. Widman, “An online growth mindset intervention in a sample of rural adolescent girls”, *British Journal of Educational Psychology* **88**, 3, 428–445 (2018).
- [57] Burstein, J., J. Tetreault and N. Madnani, “The e-rater® automated essay scoring system”, in “Handbook of automated essay evaluation”, pp. 77–89 (Routledge, 2013).
- [58] Burton, R. R., “Diagnosing bugs in a simple procedural skill”, *Intelligent Tutoring Systems* pp. 157–184 (1982).
- [59] Cahill, D. R. and R. J. Leonard, “Missteps and masquerade in american medical academe: Clinical anatomists call for action”, *Clinical Anatomy* **12**, 3, 220–222 (1999).
- [60] Callan, G. L., G. J. Marchant, W. H. Finch and R. L. German, “Metacognition, strategies, achievement, and demographics: Relationships across countries”, *Educational Sciences: Theory & Practice* **16**, 5 (2016).
- [61] Calvo, R. A., S. T. O’Rourke, J. Jones, K. Yacef and P. Reimann, “Collaborative writing support tools on the cloud”, *IEEE Transactions on Learning Technologies* **4**, 1, 88–97 (2010).
- [62] Camacho-Morles, J., G. R. Slep, R. Pekrun, K. Loderer, H. Hou and L. G. Oades, “Activity achievement emotions and academic performance: A meta-analysis”, *Educational Psychology Review* **33**, 3, 1051–1095 (2021).
- [63] Cambre, J., S. Klemmer and C. Kulkarni, “Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection”, in “Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems”, pp. 1–13 (2018).
- [64] Cardy, R. L. and G. H. Dobbins, “Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance.”, *Journal of applied psychology* **71**, 4, 672 (1986).
- [65] Carroll Massey, G., M. Vaughn Scott and S. M. Dornbusch, “Racism without racists: Institutional racism in urban schools”, *The Black Scholar* **7**, 3, 10–19 (1975).
- [66] Chan, J. Y.-C., E. R. Ottmar and J.-E. Lee, “Slow down to speed up: Longer pause time before solving problems relates to higher strategy efficiency”, *Learning and Individual Differences* **93**, 102–109 (2022).
- [67] Chan, J. Y.-C., E. R. Ottmar and J.-E. Lee, “Slow down to speed up: Longer pause time before solving problems relates to higher strategy efficiency”, *Learning and Individual Differences* **93**, 102109 (2022).
- [68] Chen, H. and B. He, “Automated essay scoring by maximizing human-machine agreement”, in “Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing”, pp. 1741–1752 (2013).
- [69] Christensen, C., M. Horn and C. Johnson (McGraw-Hill Education, New York, USA, 2017).
- [70] Cimpian, J. R., S. T. Lubienski, C. M. Ganley and Y. Copur-Gencturk, “Teachers’ perceptions of students’ mathematics proficiency may exacerbate early gender gaps in achievement.”, *Developmental psychology* **50**, 4, 1262 (2014).
- [71] Cimpian, J. R., S. T. Lubienski, J. D. Timmer, M. B. Makowski and E. K. Miller, “Have gender gaps in math closed? achievement, teacher perceptions, and learning behaviors across two ecls-k cohorts”, *AERA Open* **2**, 4, 2332858416673617 (2016).
- [72] Claro, S., D. Paunesku and C. S. Dweck, “Growth mindset tempers the effects of poverty on academic achievement”, *Proceedings of the National Academy of Sciences* **113**, 31, 8664–8668 (2016).

- [73] Clore, G. L. and J. R. Huntsinger, “How the object of affect guides its impact”, *Emotion Review* **1**, 1, 39–54 (2009).
- [74] Confrey, J., “Chapter 8: What constructivism implies for teaching”, *Journal for Research in Mathematics Education. Monograph* **4**, 107–210 (1990).
- [75] Cooney, T., W. Sanchez, K. Leatham and D. Mewborn, “Open-ended assessment in math: A searchable collection of 450+ questions”, (2004).
- [76] Cooper, W. H., “Ubiquitous halo.”, *Psychological bulletin* **90**, 2, 218 (1981).
- [77] Copur-Gencturk, Y., J. R. Cimpian, S. T. Lubienski and I. Thacker, “Teachers’ bias against the mathematical ability of female, black, and hispanic students”, *Educational Researcher* **49**, 1, 30–43 (2020).
- [78] Copur-Gencturk, Y., I. Thacker and J. R. Cimpian, “Teachers’ race and gender biases and the moderating effects of their beliefs and dispositions”, *International Journal of STEM Education* **10**, 1, 1–25 (2023).
- [79] Corbett, A. T. and J. R. Anderson, “Knowledge tracing: Modeling the acquisition of procedural knowledge”, *User modeling and user-adapted interaction* **4**, 4, 253–278 (1994).
- [80] Corbett, A. T. and J. R. Anderson, “Knowledge tracing: Modeling the acquisition of procedural knowledge”, *User modeling and user-adapted interaction* **4**, 4, 253–278 (1994).
- [81] Cox, L. S., “Diagnosing and remediating systematic errors in addition and subtraction computations.”, *Arithmetic Teacher* **22**, 2, 151–157 (1975).
- [82] Craig, S., A. Graesser, J. Sullins and B. Gholson, “Affect and learning: an exploratory look into the role of affect in learning with autotutor”, *Journal of educational media* **29**, 3, 241–250 (2004).
- [83] Crocker, J., K. Voelkl, M. Testa and B. Major, “Social stigma: The affective consequences of attributional ambiguity.”, *Journal of personality and social psychology* **60**, 2, 218 (1991).
- [84] Cronbach, L. J., *Five perspectives on the validity argument.*, p. 3–17 (Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1988).
- [85] d’Anjou, B., S. Bakker, P. An and T. Bekker, “How peripheral data visualisation systems support secondary school teachers during vle-supported lessons”, in “Proceedings of the 2019 on Designing Interactive Systems Conference”, pp. 859–870 (2019).
- [86] Davis, D., I. Jivet, R. F. Kizilcec, G. Chen, C. Hauff and G.-J. Houben, “Follow the successful crowd: raising mooc completion rates through social comparison at scale”, in “Proceedings of the seventh international learning analytics & knowledge conference”, pp. 454–463 (2017).
- [87] de Kraker-Pauw, E., F. van Wesel, T. Verwijmeren, E. Denessen and L. Krabbendam, “Are teacher beliefs gender-related?”, *Learning and Individual Differences* **51**, 333–340 (2016).
- [88] Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database”, in “2009 IEEE conference on computer vision and pattern recognition”, pp. 248–255 (Ieee, 2009).
- [89] Dennis, I., “Halo effects in grading student projects.”, *Journal of Applied Psychology* **92**, 4, 1169 (2007).
- [90] Denny, P., J. Hamer, A. Luxton-Reilly and H. Purchase, “Peerwise: students sharing their multiple choice questions”, in “Proceedings of the fourth international workshop on computing education research”, pp. 51–58 (2008).
- [91] Denny, P., B. Hanks, B. Simon and S. Bagley, “Peerwise: Exploring conflicting efficacy studies”, in “Proceedings of the seventh international workshop on Computing education research”, pp. 53–60 (2011).

- [92] Denny, P., A. Luxton-Reilly and J. Hamer, “The peerwise system of student contributed assessment questions”, in “Proceedings of the Tenth Conference on Australasian Computing Education - Volume 78”, ACE ’08, p. 69–74 (Australian Computer Society, Inc., AUS, 2008).
- [93] Denny, P., A. Luxton-Reilly and J. Hamer, “The peerwise system of student contributed assessment questions”, in “Proceedings of the tenth conference on Australasian computing education-Volume 78”, pp. 69–74 (Citeseer, 2008).
- [94] Denny, P., A. Luxton-Reilly and J. Hamer, “The peerwise system of student contributed assessment questions”, in “Proceedings of the tenth conference on Australasian computing education-Volume 78”, pp. 69–74 (Citeseer, 2008).
- [95] Denny, P., A. Luxton-Reilly and J. Hamer, “Student use of the peerwise system”, in “Proceedings of the 13th annual conference on Innovation and technology in computer science education”, pp. 73–77 (2008).
- [96] Dijkstra, P., F. X. Gibbons and A. P. Buunk, “Social comparison theory.”, *Social psychological foundations of clinical psychology* (2010).
- [97] Dikli, S., “An overview of automated scoring of essays”, *The Journal of Technology, Learning and Assessment* **5**, 1 (2006).
- [98] Dillenbourg, P. and P. Jermann, “Technology for classroom orchestration”, in “New science of learning”, pp. 525–552 (Springer, 2010).
- [99] Dillenbourg, P., G. Zufferey, H. Alavi, P. Jermann, S. Do-Lenh, Q. Bonnard, S. Cuendet and F. Kaplan, “Classroom orchestration: The third circle of usability”, (2011).
- [100] D’Mello, S., E. Dieterle and A. Duckworth, “Advanced, analytic, automated (aaa) measurement of engagement during learning”, *Educational psychologist* **52**, 2, 104–123 (2017).
- [101] Doroudi, S., “Mastery learning heuristics and their hidden models”, in “Artificial Intelligence in Education”, vol. 12164, p. 86–91 (Springer International Publishing, Morocco, 2020).
- [102] Doroudi, S. and E. Brunskill, “Fairer but not fair enough on the equitability of knowledge tracing”, in “Proceedings of the 9th International Conference on Learning Analytics and Knowledge”, LAK19, p. 335–339 (Association for Computing Machinery, New York, NY, USA, 2019), URL <https://doi.org/10.1145/3303772.3303838>.
- [103] Doroudi, S., J. Williams, J. Kim, T. Patikorn, K. Ostrow, D. Selent, N. T. Heffernan, T. Hills and C. Rosé, “Crowdsourcing and education: Towards a theory and praxis of learnersourcing”, (International Society of the Learning Sciences, Inc.[ISLS]., 2018).
- [104] Dougiamas, M. and P. C. Taylor, “Interpretive analysis of an internet-based course constructed using a new courseware tool called moodle”, in “2nd conference of herdsa (the higher education research and development society of australasia)”, pp. 7–10 (HERDSA, Perth Western Australia, 2002).
- [105] Duckworth, A. L., C. Peterson, M. D. Matthews and D. R. Kelly, “Grit: Perseverance and passion for long-term goals”, *Journal of Personality and Social Psychology* **92**, 6, 1087–1101 (2007).
- [106] Dweck, C. S., *Mindset: The new psychology of success* (Random House, 2006).
- [107] Dzhafarov, E. N. and R. Schweickert, “Decompositions of response times: An almost general theory”, *Journal of Mathematical Psychology* **39**, 3, 285–314 (1995).
- [108] D’Mello, S., B. Lehman, R. Pekrun and A. Graesser, “Confusion can be beneficial for learning”, *Learning and Instruction* **29**, 153–170 (2014).
- [109] D’Mello, S., B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins and A. Graesser, “A time for emoting: When affect-sensitivity is and isn’t effective at promoting deep learning”, in “International conference on intelligent tutoring systems”, pp. 245–254 (Springer, 2010).

- [110] Ebert, D., “Graphing projects with desmos”, *The Mathematics Teacher* **108**, 5, 388–391 (2014).
- [111] Elias, S., “Seethal/sentiment_analysis_generic_dataset”, URL https://huggingface.co/Seethal/sentiment_analysis_generic_dataset (2022).
- [112] Ellis, H. C. and P. W. Ashbrook, “The” state” of mood and memory research: A selective review”, *Journal of Social Behavior and Personality* **4**, 2, 1 (1989).
- [113] Erickson, J. A., A. F. Botelho, S. McAteer, A. Varatharaj and N. T. Heffernan, “The automated grading of student open responses in mathematics”, in “Proceedings of the Tenth International Conference on Learning Analytics & Knowledge”, pp. 615–624 (2020).
- [114] Espinoza, P., A. B. Arêas da Luz Fontes and C. J. Arms-Chavez, “Attributional gender bias: Teachers’ ability and effort explanations for students’ math performance”, *Social Psychology of Education* **17**, 105–126 (2014).
- [115] Fajardo, D. M., “Author race, essay quality, and reverse discrimination”, *Journal of Applied Social Psychology* **15**, 3, 255–268 (1985).
- [116] Falch, T. and L. R. Naper, “Educational evaluation schemes and gender gaps in student achievement”, *Economics of Education Review* **36**, 12–25 (2013).
- [117] Feldman, R., “Techniques and applications for sentiment analysis”, *Communications of the ACM* **56**, 4, 82–89 (2013).
- [118] Feng, M. and N. T. Heffernan, “Informing teachers live about student learning: Reporting in the assistment system”, *Technology Instruction Cognition and Learning* **3**, 1/2, 63 (2006).
- [119] Ferreira, J. and A. Zwinderman, “On the benjamini–hochberg method”, *The Annals of Statistics* **34**, 4, 1827–1849 (2006).
- [120] Ferren, M., “Remote learning and school reopenings: What worked and what didn’t”, URL <https://www.americanprogress.org/article/remote-learning-school-reopenings-worked-didnt/> (2021).
- [121] Flavell, J. H., “Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry.”, *American psychologist* **34**, 10, 906 (1979).
- [122] for Science, N. C. and E. Statistics., “Women, minorities, and persons with disabilities in science and engineering 2023. arlington, va.”, Tech. rep., URL <https://nces.nsf.gov/pubs/nsf23315>, accessed 06-17-2022 (2023).
- [123] Forgas, J. P., “She just doesn’t look like a philosopher...? affective influences on the halo effect in impression formation”, *European Journal of Social Psychology* **41**, 7, 812–817 (2011).
- [124] Forum, W. E., “The rise of online learning during the covid-19 pandemic — world economic forum”, URL <https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/> (2020).
- [125] Foundation., N. S., “Science and engineering degrees, by race/ethnicity of recipients: 2008–18. arlington, va.”, Tech. rep., URL <https://ncesdata.nsf.gov/sere/2018/>, accessed 06-17-2022 (2020).
- [126] Fowler, M., B. Chen, S. Azad, M. West and C. Zilles, “Autograding” explain in plain english” questions using nlp”, in “Proceedings of the 52nd ACM Technical Symposium on Computer Science Education”, pp. 1163–1169 (2021).
- [127] Franz, W.-J. I., “Grade inflation under the threat of students’ nuisance: Theory and evidence”, *Economics of Education Review* **29**, 3, 411–422 (2010).
- [128] Funk, S. C. and K. L. Dickson, “Multiple-choice and short-answer exam performance in a college classroom”, *Teaching of Psychology* **38**, 4, 273–277 (2011).

- [129] Gerber, J., L. Wheeler and J. Suls, “A social comparison theory meta-analysis 60+ years on.”, *Psychological bulletin* **144**, 2, 177 (2018).
- [130] Gibbons, S. and A. Chevalier, “Assessment and age 16+ education participation”, *Research Papers in Education* **23**, 2, 113–123 (2008).
- [131] Gold, C. and T. Zesch, “Exploring the impact of handwriting recognition on the automated scoring of handwritten student answers”, in “2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)”, pp. 252–257 (IEEE, 2020).
- [132] Gough, E., D. DeJong, T. Grundmeyer and M. Baron, “K-12 teacher perceptions regarding the flipped classroom model for teaching and learning”, *Journal of Educational Technology Systems* **45**, 3, 390–423 (2017).
- [133] Gould, E. S., R. Li and S. J. Southard, “Effectiveness of hints vs. complete explanation using assistments”, (2010).
- [134] Graham, S., “Communicating sympathy and anger to black and white children: The cognitive (attributional) consequences of affective cues.”, *Journal of Personality and Social Psychology* **47**, 1, 40 (1984).
- [135] Graham, S. and D. Perin, “Writing next-effective strategies to improve writing of adolescents in middle and high schools”, (2007).
- [136] Gray, L. and L. Lewis, “Use of educational technology for instruction in public schools: 2019–20.”, URL <https://nces.ed.gov/pubs2021/2021017Summary.pdf> (2021).
- [137] Green, L. S., F. A. Inan and N. J. Maushak, “A case study: The role of student-generated vidcasts in k–12 language learner academic language and content acquisition”, *Journal of research on technology in education* **46**, 3, 297–324 (2014).
- [138] Greenwald, A. G., D. E. McGhee and J. L. Schwartz, “Measuring individual differences in implicit cognition: the implicit association test.”, *Journal of personality and social psychology* **74**, 6, 1464 (1998).
- [139] Große, C. S. and A. Renkl, “Finding and fixing errors in worked examples: Can this foster learning outcomes?”, *Learning and instruction* **17**, 6, 612–634 (2007).
- [140] Guo, R., D. Palmer-Brown, S. W. Lee and F. F. Cai, “Intelligent diagnostic feedback for online multiple-choice questions”, *Artificial Intelligence Review* **42**, 3, 369–383 (2014).
- [141] Gurung, A., S. Baral, K. P. Vanacore, A. A. Mcreynolds, H. Kreisberg, A. F. Botelho, S. T. Shaw and N. T. Hefferna, “Identification, exploration, and remediation: Can teachers predict common wrong answers?”, in “LAK23: 13th International Learning Analytics and Knowledge Conference”, pp. 399–410 (2023).
- [142] Gurung, A., A. F. Botelho and N. T. Heffernan, “Examining student effort on help through response time decomposition”, in “LAK21: 11th International Learning Analytics and Knowledge Conference”, pp. 292–301 (2021).
- [143] Gurung, A., A. F. Botelho and N. T. Heffernan, “Examining student effort on help through response time decomposition”, in “LAK21: 11th International Learning Analytics and Knowledge Conference”, LAK21, pp. 292—301 (Association for Computing Machinery, New York, NY, USA, 2021), URL <https://doi.org/10.1145/3448139.3448167>.
- [144] Gurung, A., A. F. Botelho, R. Thompson, A. C. Sales, S. Baral and N. T. Heffernan, “Considerate, unfair, or just fatigued? examining factors that impact teacher”, in “Proceedings of the 30th International Conference on Computers in Education.”, (2022).
- [145] Guskey, T. R. and T. D. Pigott, “Research on group-based mastery learning programs: A meta-analysis”, *The Journal of Educational Research* **81**, 4, 197–216 (1988).

- [146] Gutman, L. M. and I. Schoon, “The impact of non-cognitive skills on outcomes for young people. a literature review”, (2013).
- [147] Harber, K. D., “The positive feedback bias as a response to out-group unfriendliness 1”, *Journal of Applied Social Psychology* **34**, 11, 2272–2297 (2004).
- [148] Harber, K. D., J. L. Gorman, F. P. Gengaro, S. Butisingh, W. Tsang and R. Ouellette, “Students’ race and teachers’ social support affect the positive feedback bias in public schools.”, *Journal of Educational Psychology* **104**, 4, 1149 (2012).
- [149] Harber, K. D., R. Stafford and K. A. Kennedy, “The positive feedback bias as a response to self-image threat”, *British Journal of Social Psychology* **49**, 1, 207–218 (2010).
- [150] Harrison, C. J., K. D. Könings, L. W. T. Schuwirth, V. Wass and C. P. M. van der Vleuten, “Changing the culture of assessment: the dominance of the summative assessment paradigm”, *BMC Medical Education* **17**, 1, 73 (2017).
- [151] Hasni, A., F. Bousadra, V. Belletête, A. Benabdallah, M.-C. Nicole and N. Dumais, “Trends in research on project-based science and technology teaching and learning at k–12 levels: a systematic review”, *Studies in Science Education* **52**, 2, 199–231 (2016).
- [152] Hattie, J., “Influences on student learning”, Inaugural lecture given on August 2, 1999, 21 (1999).
- [153] Hattie, J. and H. Timperley, “The power of feedback”, *Review of educational research* **77**, 1, 81–112 (2007).
- [154] Heffernan, N. T. and C. L. Heffernan, “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, 470–497 (2014).
- [155] Heffernan, N. T. and C. L. Heffernan, “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, 470–497 (2014).
- [156] Heiner, C., J. Beck and J. Mostow, “Improving the help selection policy in a reading tutor that listens”, in “InSTIL/ICALL Symposium 2004”, (2004).
- [157] Henry, P. and H. Semple, “Integrating online gis into the k–12 curricula: Lessons from the development of a collaborative gis in michigan”, *Journal of Geography* **111**, 1, 3–14 (2012).
- [158] Hicks, C. M., V. Pandey, C. A. Fraser and S. Klemmer, “Framing feedback: Choosing review environment features that support high quality peer assessment”, in “Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems”, CHI ’16, p. 458–469 (Association for Computing Machinery, New York, NY, USA, 2016), URL <https://doi.org/10.1145/2858036.2858195>.
- [159] Hill, H. C., D. L. Ball and S. G. Schilling, “Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers’ topic-specific knowledge of students”, *Journal for research in mathematics education* **39**, 4, 372–400 (2008).
- [160] Hill, H. C. and M. Chin, “Connections between teachers’ knowledge of students, instruction, and achievement outcomes”, *American Educational Research Journal* **55**, 5, 1076–1112 (2018).
- [161] Hill, H. C., S. G. Schilling and D. L. Ball, “Developing measures of teachers’ mathematics knowledge for teaching”, *The elementary school journal* **105**, 1, 11–30 (2004).
- [162] Hills, T. T., “Crowdsourcing content creation in the classroom”, *Journal of Computing in Higher Education* **27**, 1, 47–67 (2015).
- [163] Hinnerich, B. T., E. Höglin and M. Johannesson, “Are boys discriminated in swedish high schools?”, *Economics of Education review* **30**, 4, 682–690 (2011).
- [164] Hochanadel, A. and D. Finamore, “Fixed and growth mindset in education and how grit helps students persist in the face of adversity”, *Journal of International Education Research* **11**, 1, 47–50 (2015).

- [165] Hohenwarter, M. and M. Hohenwarter, “Geogebra”, Available on-line at <http://www.geogebra.org/cms/en> (2002).
- [166] Holstein, K. and S. Doroudi, “Equity and artificial intelligence in education: Will “aied” amplify or alleviate inequities in education?”, URL <https://arxiv.org/abs/2104.12920> (2021).
- [167] Holstein, K., B. M. McLaren and V. Alevan, “Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms”, in “International conference on artificial intelligence in education”, pp. 154–168 (Springer, 2018).
- [168] Holstein, K., B. M. McLaren and V. Alevan, “Co-designing a real-time classroom orchestration tool to support teacher–ai complementarity”, *Journal of Learning Analytics* **6**, 2 (2019).
- [169] House, W. W. C., “Procedures and standards handbook (v5).”, URL <https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-HandbookVer5.0AppIES-508.pdf> (2022).
- [170] Hu, D., “How khan academy is using machine learning to assess student mastery”, URL <http://david-hu.com/2011/11/02/how-khan-academy-is-using-machine-learning-to-assess-student-mastery.html> (2011).
- [171] Hu, S., A. C. McCormick and R. M. Gonyea, “Examining the relationship between student learning and persistence”, *Innovative Higher Education* **37**, 5, 387–395 (2012).
- [172] Huff, J. D. and J. L. Nietfeld, “Using strategy instruction and confidence judgments to improve metacognitive monitoring”, *Metacognition and Learning* **4**, 2, 161–176 (2009).
- [173] Inc., G., “Education technology use in schools.”, URL <https://www.newschools.org/wp-content/uploads/2020/03/Gallup-Ed-Tech-Use-in-Schools-2.pdf> (2019).
- [174] Inc., I., “Canvas”, URL <https://www.instructure.com/canvas/resources/research> (2019).
- [175] Izurieta, C. and J. M. Bieman, “A multiple case study of design pattern decay, grime, and rot in evolving software systems”, *Software Quality Journal* **21**, 2, 289–323 (2013).
- [176] Jacob, R., H. Hill and D. Corey, “The impact of a professional development program on teachers’ mathematical knowledge for teaching, instruction, and student achievement”, *Journal of Research on Educational Effectiveness* **10**, 2, 379–407 (2017).
- [177] Joyner, D. A., W. Ashby, L. Irish, Y. Lam, J. Langston, I. Lupiani, M. Lustig, P. Pettoruto, D. Sheahen, A. Smiley, A. Bruckman and A. Goel, “Graders as meta-reviewers: Simultaneously scaling and improving expert evaluation for large online classrooms”, in “Proceedings of the Third (2016) ACM Conference on Learning @ Scale”, L@S ’16, p. 399–408 (Association for Computing Machinery, New York, NY, USA, 2016), URL <https://doi.org/10.1145/2876034.2876044>.
- [178] Jurman, G., S. Riccadonna, R. Visintainer and C. Furlanello, “Canberra distance on ranked lists”, in “Proceedings of advances in ranking NIPS 09 workshop”, pp. 22–27 (Citeseer, 2009).
- [179] Kelly, K., Y. Wang, T. Thompson and N. Heffernan, “Defining mastery: Knowledge tracing versus n-consecutive correct responses”, *Student Modeling From Different Aspects* p. 39 (2016).
- [180] Kelly, K. M., Y. Wang, T. Thompson and N. T. Heffernan, “Defining mastery: Knowledge tracing versus n-consecutive correct responses”, in “proceedings of the 8th International Conference on Educational Data Mining”, pp. 39–46 (Association for Computing Machinery, New York, NY, USA, 2015), URL <http://web.wpi.edu/Pubs/ETD/Available/etd-041416-122623/unrestricted/wang.pdf#page=42>.
- [181] Khajah, M., R. V. Lindsey and M. C. Mozer, “How deep is knowledge tracing?”, *CoRR abs/1604.02416*, URL <http://arxiv.org/abs/1604.02416> (2016).

- [182] Kharrufa, A., S. Rix, T. Osadchiy, A. Preston and P. Olivier, “Group spinner: recognizing and visualizing learning in the classroom for reflection, communication, and planning”, in “Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems”, pp. 5556–5567 (2017).
- [183] Khine, M. S. and S. Areepattamannil, *Non-cognitive skills and factors in educational attainment* (Springer, 2016).
- [184] Khosravi, H., K. Kitto and J. J. Williams, “Ripple: A crowdsourced adaptive platform for recommendation of learning activities”, arXiv preprint arXiv:1910.05522 (2019).
- [185] Khuong, V. T. M., H. Q. Ung, C. T. Nguyen and M. Nakagawa, “Clustering offline handwritten mathematical answers for computer-assisted marking”, in “Proc. 1st Int. Conf. on Pattern Recognit. and Artificial Intelligence, Montreal, Canada”, pp. 121–126 (2018).
- [186] Kieran, C., “Concepts associated with the equality symbol”, *Educational studies in Mathematics* **12**, 317–326 (1981).
- [187] Kim, J. *et al.*, *Learnersourcing: improving learning with collective learner activity*, Ph.D. thesis, Massachusetts Institute of Technology (2015).
- [188] Kim, Y.-S. G., C. Schatschneider, J. Wanzek, B. Gatlin and S. Al Otaiba, “Writing evaluation: rater and task effects on the reliability of writing scores for children in grades 3 and 4”, *Reading and writing* **30**, 6, 1287–1310 (2017).
- [189] King, R. B. and J. E. Trinidad, “Growth mindset predicts achievement only among rich students: examining the interplay between mindset and socioeconomic status”, *Social Psychology of Education* **24**, 3, 635–652 (2021).
- [190] Kizilcec, R. F., J. Reich, M. Yeomans, C. Dann, E. Brunskill, G. Lopez, S. Turkay, J. J. Williams and D. Tingley, “Scaling up behavioral science interventions in online education”, *Proceedings of the National Academy of Sciences* **117**, 26, 14900–14905 (2020).
- [191] Klufa, J., “Multiple choice question tests – advantages and disadvantages”, *Mathematics and Computers in Sciences and Industry* pp. 91–97 (2015).
- [192] Kluger, A. N. and A. DeNisi, “The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory.”, *Psychological bulletin* **119**, 2, 254 (1996).
- [193] Koedinger, K. R., J. R. Anderson, W. H. Hadley, M. A. Mark *et al.*, “Intelligent tutoring goes to school in the big city”, *International Journal of Artificial Intelligence in Education* **8**, 1, 30–43 (1997).
- [194] Kohavi, R., R. Longbotham, D. Sommerfield and R. M. Henne, “Controlled experiments on the web: survey and practical guide”, *Data Mining and Knowledge Discovery* **18**, 1, 140–181 (2009).
- [195] Kramarski, B. and M. Gutman, “How can self-regulated learning be supported in mathematical e-learning environments?”, *Journal of Computer Assisted Learning* **22**, 1, 24–33 (2006).
- [196] Kulik, C.-L. C., J. A. Kulik and R. L. Bangert-Drowns, “Effectiveness of mastery learning programs: A meta-analysis”, *Review of Educational Research* **60**, 2, 265–299 (1990).
- [197] Kulkarni, C. E., M. S. Bernstein and S. R. Klemmer, “Peerstudio: Rapid peer feedback emphasizes revision and improves performance”, in “Proceedings of the Second (2015) ACM Conference on Learning @ Scale”, L@S ’15, p. 75–84 (Association for Computing Machinery, New York, NY, USA, 2015), URL <https://doi.org/10.1145/2724660.2724670>.
- [198] Lai, C.-F., “Error analysis in mathematics. technical report# 1012.”, *Behavioral Research and Teaching* (2012).
- [199] Lan, A. S., D. Vats, A. E. Waters and R. G. Baraniuk, “Mathematical language processing: Automatic grading and feedback for open response mathematical questions”, in “Proceedings of the second (2015) ACM conference on learning@ scale”, pp. 167–176 (2015).

- [200] Landy, D. and H. Sigall, "Beauty is talent: Task evaluation as a function of the performer's physical attractiveness.", *Journal of Personality and Social Psychology* **29**, 3, 299 (1974).
- [201] Lavy, V., "Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment", *Journal of public Economics* **92**, 10-11, 2083–2105 (2008).
- [202] Lawson, A. P., A. Mirinjian and J. Y. Son, "Can preventing calculations help students learn math?", *Journal of Cognitive Education and Psychology* **17**, 2, 178–197 (2019).
- [203] Lebuda, I. and M. Karwowski, "Tell me your name and i'll tell you how creative your work is: Author's name and gender as factors influencing assessment of products' creativity in four different domains", *Creativity Research Journal* **25**, 1, 137–142 (2013).
- [204] Lee, J. and L. Stankov, "Higher-order structure of motivation, self-beliefs, learning strategies, and attitudes toward school and its prediction of pisa 2003 mathematics scores", *Learning and Individual Differences* **26**, 119–130 (2013).
- [205] Leelawong, K. and G. Biswas, "Designing learning by teaching agents: The betty's brain system", *International Journal of Artificial Intelligence in Education* **18**, 3, 181–208 (2008).
- [206] Lehman, B., M. Matthews, S. D'Mello and N. Person, "What are you feeling? investigating student affective states during expert human tutoring sessions", in "International conference on intelligent tutoring systems", pp. 50–59 (Springer, 2008).
- [207] Leibold, N. and L. M. Schwarz, "The art of giving online feedback.", *Journal of Effective Teaching* **15**, 1, 34–46 (2015).
- [208] Levin, N., R. S. Baker, N. Nasiar, S. Fancsali and S. Hutt, "Evaluating gaming detector model robustness over time", in "Proceedings of the 15th International Conference on Educational Data Mining, International Educational Data Mining Society", (2022).
- [209] Levy, Y., "Comparing dropouts and persistence in e-learning courses", *Computers in Education* **48**, 2, 185–204 (2007).
- [210] Li, J., "Learning vocabulary via computer-assisted scaffolding for text processing", *Computer Assisted Language Learning* **23**, 3, 253–275 (2010).
- [211] Limeri, L. B., J. Choe, H. G. Harper, H. R. Martin, A. Benton and E. L. Dolan, "Knowledge or abilities? how undergraduates define intelligence", *CBE—Life Sciences Education* **19**, 1, ar5 (2020).
- [212] Linnenbrink, E. A., "The role of affect in student learning: A multi-dimensional approach to considering the interaction of affect, motivation, and engagement", in "Emotion in education", pp. 107–124 (Elsevier, 2007).
- [213] Little, J. L. and E. L. Bjork, "Optimizing multiple-choice tests as tools for learning", *Memory & Cognition* **43**, 1, 14–26 (2015).
- [214] Little, J. L., E. L. Bjork, R. A. Bjork and G. Angello, "Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting", *Psychological science* **23**, 11, 1337–1344 (2012).
- [215] Livne, N. L., O. E. Livne and C. A. Wight, "Enhancing mathematical creativity through multiple solution to open-ended problems online", Diperoleh dari http://www.iste.org/Content/NavigationMenu/Research/NECC_Research_Paper_Archives/NECC2008/Livne.pdf (2008).
- [216] Long, H., "What we think we know about self-directed learning", *Practice and theory in self-directed learning* pp. 1–14 (2000).
- [217] Lysakowski, R. S. and H. J. Walberg, "Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis", *American Educational Research Journal* **19**, 4, 559–572 (1982).

- [218] M. DiCosola III, B. and G. Neff, “Nudging behavior change: Using in-group and out-group social comparisons to encourage healthier choices”, in “CHI Conference on Human Factors in Computing Systems”, pp. 1–14 (2022).
- [219] Magliano, J. P., K. Millis, Y. Ozuru and D. S. McNamara, “A multidimensional framework to evaluate reading assessment tools”, in “Reading Comprehension Strategies: Theories, Interventions, and Technologies”, edited by D. S. McNamara, vol. 1, chap. A Multidimensional Framework to Evaluate Reading Assessment Tools, pp. 107–136 (Psychology Press, New York, USA, 2007), 1 edn.
- [220] Malouff, J. M., A. J. Emmerton and N. S. Schutte, “The risk of a halo bias as a reason to keep students anonymous during grading”, *Teaching of Psychology* **40**, 3, 233–237 (2013).
- [221] Malouff, J. M., S. J. Stein, L. N. Bothma, K. Coulter and A. J. Emmerton, “Preventing halo bias in grading the work of university students”, *Cogent Psychology* **1**, 1, 988937 (2014).
- [222] Marsh, H. W., U. Trautwein, O. Lüdtke and O. Köller, “Social comparison and big-fish-little-pond effects on self-concept and other self-belief constructs: Role of generalized and specific others.”, *Journal of Educational Psychology* **100**, 3, 510 (2008).
- [223] Marsh, H. W., U. Trautwein, O. Lüdtke and O. Köller, “Social comparison and big-fish-little-pond effects on self-concept and other self-belief constructs: Role of generalized and specific others.”, *Journal of Educational Psychology* **100**, 3, 510 (2008).
- [224] Martin, W. D., “The sex factor in grading composition”, *Research in the Teaching of English* **6**, 1, 36–47 (1972).
- [225] Martinez-Maldonado, R., A. Clayphan and J. Kay, “Deploying and visualising teacher’s scripts of small group activities in a multi-surface classroom ecology: A study in-the-wild”, *Computer Supported Cooperative Work (CSCW)* **24**, 2-3, 177–221 (2015).
- [226] Martinez-Maldonado, R., A. Pardo, N. Mirriahi, K. Yacef, J. Kay and A. Clayphan, “The latux workflow: designing and deploying awareness tools in technology-enabled learning settings”, in “Proceedings of the Fifth International Conference on Learning Analytics and Knowledge”, pp. 1–10 (2015).
- [227] Mavrikis, M., S. Gutierrez-Santos and A. Poulouvasilis, “Design and evaluation of teacher assistance tools for exploratory learning environments”, in “Proceedings of the Sixth International Conference on Learning Analytics & Knowledge”, pp. 168–172 (2016).
- [228] McGee, E. O. and D. B. Martin, ““you would not believe what i have to go through to prove my intellectual value!” stereotype management among academically successful black mathematics and engineering students”, *American Educational Research Journal* **48**, 6, 1347–1389 (2011).
- [229] McHugh, M. L., “Interrater reliability: the kappa statistic”, *Biochemia medica* **22**, 3, 276–282 (2012).
- [230] Meinhardt, J. and R. Pekrun, “Attentional resource allocation to emotional events: An erp study”, *Cognition and Emotion* **17**, 3, 477–500 (2003).
- [231] Mengelkamp, C. and M. Bannert, *Confidence Judgments in Learning*, p. 756–759 (Springer US, Boston, MA, 2012), URL https://doi.org/10.1007/978-1-4419-1428-6_1726.
- [232] Merriam, S. B. and L. M. Baumgartner, *Learning in adulthood: A comprehensive guide* (John Wiley & Sons, 2020).
- [233] Milkman, K. L., L. Gandhi, M. S. Patel, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, J. Rothschild, J. E. Bogard *et al.*, “A 680,000-person megastudy of nudges to encourage vaccination in pharmacies”, *Proceedings of the National Academy of Sciences* **119**, 6, e2115126119 (2022).
- [234] Mills, C., S. D’Mello, N. Bosch and A. M. Olney, “Mind wandering during learning with an intelligent tutoring system”, in “International conference on artificial intelligence in education”, pp. 267–276 (Springer, 2015).

- [235] Mitrovic, A., “An intelligent sql tutor on the web”, *International Journal of Artificial Intelligence in Education* **13**, 2-4, 173–197 (2003).
- [236] Mo, D., L. Zhang, J. Wang, W. Huang, Y. Shi, M. Boswell and S. Rozelle, “Persistence of learning gains from computer assisted learning: Experimental evidence from china.”, *Journal of Computer Assisted Learning* **31**, 6, 562–581 (2015).
- [237] Mogessie, M., J. E. Richey, B. M. McLaren, J. M. L. Andres-Bray and R. S. Baker, “Confrustion and gaming while learning with erroneous examples in a decimals game”, in “International Conference on Artificial Intelligence in Education”, pp. 208–213 (Springer, 2020).
- [238] Mohammadhassan, N., A. Mitrovic and K. Neshatian, “Investigating the effect of nudges for improving comment quality in active video watching”, *Computers & Education* **176**, 104340 (2022).
- [239] Möller, J., S. Zitzmann, F. Helm, N. Machts and F. Wolff, “A meta-analysis of relations between achievement and self-concept”, *Review of Educational Research* **90**, 3, 376–419 (2020).
- [240] Monkaresi, H., N. Bosch, R. A. Calvo and S. K. D’Mello, “Automated detection of engagement using video-based estimation of facial expressions and heart rate”, *IEEE Transactions on Affective Computing* **8**, 1, 15–28 (2016).
- [241] MOORE, S., H. NGUYEN and J. STAMPER, “Utilizing crowdsourcing and topic modeling to generate knowledge components for math and writing problems”, in “Proceedings of the 28th International Conference on Computers in Education”, pp. 31–40 (2020).
- [242] Moraveji, N., M. Morris, D. Morris, M. Czerwinski and N. H. Riche, “Classsearch: facilitating the development of web search skills through social learning”, in “Proceedings of the SIGCHI Conference on Human Factors in Computing Systems”, pp. 1797–1806 (2011).
- [243] Morris, S. S., *The relationship between self-directed learning readiness and academic performance in a nontraditional higher education program*, Ph.D. thesis, The University of Oklahoma (1995).
- [244] Mottus, A., S. Graf, N.-S. Chen *et al.*, “Use of dashboards and visualization techniques to support teacher decision making”, in “Ubiquitous learning environments and technologies”, pp. 181–199 (Springer, 2015).
- [245] Mueller, C. M. and C. S. Dweck, “Praise for intelligence can undermine children’s motivation and performance.”, *Journal of personality and social psychology* **75**, 1, 33 (1998).
- [246] Munzner, T., “A nested model for visualization design and validation”, *IEEE transactions on visualization and computer graphics* **15**, 6, 921–928 (2009).
- [247] Murphy, R., J. Roschelle, M. Feng and C. A. Mason, “Investigating efficacy, moderators and mediators for an online mathematics homework intervention”, *Journal of Research on Educational Effectiveness* **13**, 2, 235–270 (2020).
- [248] Myers, E. and M. Souza, “Social comparison nudges without monetary incentives: Evidence from home energy reports”, *Journal of Environmental Economics and Management* **101**, 102315 (2020).
- [249] Narciss, S., “The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning.”, *Experimental psychology* **51**, 3, 214 (2004).
- [250] Narciss, S., “Designing and evaluating tutoring feedback strategies for digital learning”, *Digital Education Review* , 23, 7–26 (2013).
- [251] National Governors Association Center for Best Practices, C. o. C. S. S. O., “Common core state standards (mathematics standards)”, URL <http://www.corestandards.org/Math/> (2010).
- [252] Nieva, V. F. and B. A. Gutek, “Sex effects on evaluation”, *Academy of management Review* **5**, 2, 267–276 (1980).
- [253] Nisbett, R. E. and T. D. Wilson, “The halo effect: Evidence for unconscious alteration of judgments.”, *Journal of personality and social psychology* **35**, 4, 250 (1977).

- [254] Nürnberger, M., J. Nerb, F. Schmitz, J. Keller and S. Sütterlin, “Implicit gender stereotypes and essentialist beliefs predict preservice teachers’ tracking recommendations”, *The Journal of Experimental Education* **84**, 1, 152–174 (2016).
- [255] of Education Sciences; National Center for Education Evaluation, I. and R. Assistance, “Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide”, URL <http://doi.apa.org/get-pe-doi.cfm?doi=10.1037/e370412004-001> (2003).
- [256] of Students, N. U., “Higher education campaign: Mark my words, not my name”, Tech. rep., NUS, URL <http://samairaanjum.weebly.com/uploads/1/0/5/2/10526755/markmywordsbrief1-1.pdf>, accessed 06-17-2022 (2008).
- [257] Ojose, B., *Common misconceptions in mathematics: Strategies to correct them* (University Press of America, 2015).
- [258] Ojose, B., “Students’ misconceptions in mathematics: Analysis of remedies and what research says.”, *Ohio Journal of School Mathematics* , 72 (2015).
- [259] Onu, V., M. Eskay, J. Igbo, N. Obiyo and O. Agbo, “Effect of training in math metacognitive strategy on fractional achievement of nigerian schoolchildren.”, Online Submission (2012).
- [260] O’Rourke, E., K. Haimovitz, C. Ballweber, C. Dweck and Z. Popović, “Brain points: A growth mindset incentive structure boosts persistence in an educational game”, in “Proceedings of the SIGCHI conference on human factors in computing systems”, pp. 3339–3348 (2014).
- [261] Ostrow, K. S. and N. T. Heffernan, “The role of student choice within adaptive tutoring”, in “International Conference on Artificial Intelligence in Education”, pp. 752–755 (Springer, 2015).
- [262] Pandey, S. and G. Karypis, “A self-attentive model for knowledge tracing”, arXiv preprint arXiv:1907.06837 (2019).
- [263] Pane, J. F., B. A. Griffin, D. F. McCaffrey and R. Karam, “Effectiveness of cognitive tutor algebra i at scale”, *Educational Evaluation and Policy Analysis* **36**, 2, 127–144 (2014).
- [264] Pane, J. F., D. F. McCaffrey, M. E. Slaughter, J. L. Steele and G. S. Ikemoto, “An experiment to evaluate the efficacy of cognitive tutor geometry”, *Journal of Research on Educational Effectiveness* **3**, 3, 254–281 (2010).
- [265] Paquette, L. and R. S. Baker, “Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system”, *Interactive Learning Environments* **27**, 5-6, 585–597 (2019).
- [266] Paquette, L., R. S. Baker, A. de Carvalho and J. Ocumpaugh, “Cross-system transfer of machine learned and knowledge engineered models of gaming the system”, in “International Conference on User Modeling, Adaptation, and Personalization”, pp. 183–194 (Springer, 2015).
- [267] Paquette, L., A. de Carvahlo, R. Baker and J. Ocumpaugh, “Reengineering the feature distillation process: A case study in detection of gaming the system”, in “Educational data mining 2014”, (Citeseer, 2014).
- [268] Pardos, Z., S. Farrar, J. Kolb, G. X. Peh and J. H. Lee, “Distributed representation of misconceptions”, (International Society of the Learning Sciences, Inc.[ISLS]., 2018).
- [269] Pardos, Z. A., R. S. Baker, M. O. San Pedro, S. M. Gowda and S. M. Gowda, “Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes”, in “Proceedings of the third international conference on learning analytics and knowledge”, pp. 117–124 (2013).
- [270] Pardos, Z. A. and N. T. Heffernan, “Modeling individualization in a bayesian networks implementation of knowledge tracing”, in “International Conference on User Modeling, Adaptation, and Personalization”, pp. 255–266 (Springer, 2010).

- [271] Pardos, Z. A. and N. T. Heffernan, “Kt-idem: Introducing item difficulty to the knowledge tracing model”, in “International conference on user modeling, adaptation, and personalization”, pp. 243–254 (Springer, 2011).
- [272] Park, D., E. Tsukayama, A. Yu and A. L. Duckworth, “The development of grit and growth mindset during adolescence”, *Journal of Experimental Child Psychology* **198**, 104889 (2020).
- [273] Patikorn, T. and N. T. Heffernan, “Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms”, in “Proceedings of the Seventh ACM Conference on Learning@Scale”, pp. 115–124 (2020).
- [274] Patikorn, T. and N. T. Heffernan, “Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms”, in “Proceedings of the Seventh ACM Conference on Learning@Scale”, pp. 115–124 (2020).
- [275] Pavlik Jr, P. I., H. Cen and K. R. Koedinger, “Performance factors analysis—a new alternative to knowledge tracing.”, Online Submission (2009).
- [276] Pekrun, R., “The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice”, *Educational psychology review* **18**, 4, 315–341 (2006).
- [277] Pekrun, R., T. Goetz, W. Titz and R. P. Perry, “Academic emotions in students’ self-regulated learning and achievement: A program of qualitative and quantitative research”, *Educational psychologist* **37**, 2, 91–105 (2002).
- [278] Picciano, A. G., J. Seaman, P. Shea and K. Swan, “Examining the extent and nature of online learning in american k-12 education: The research initiatives of the alfred p. sloan foundation”, *The internet and higher education* **15**, 2, 127–135 (2012).
- [279] Piché, G. L., M. Michlin, D. Rubin and A. Sullivan, “Effects of dialect-ethnicity, social class and quality of written compositions on teachers’ subjective evaluations of children”, *Communications Monographs* **44**, 1, 60–72 (1977).
- [280] Piech, C., J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas and J. Sohl-Dickstein, “Deep knowledge tracing”, *Advances in neural information processing systems* **28** (2015).
- [281] Piech, C., J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas and J. Sohl-Dickstein, “Deep knowledge tracing”, in “Advances in Neural Information Processing Systems”, edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, vol. 28 (Curran Associates, Inc., 2015), URL <https://proceedings.neurips.cc/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf>.
- [282] Pitt, E. and N. Winstone, “The impact of anonymous marking on students’ perceptions of fairness, feedback and relationships with lecturers”, *Assessment & Evaluation in Higher Education* **43**, 7, 1183–1193 (2018).
- [283] Plak, S., C. van Klaveren and I. Cornelisz, “Raising student engagement using digital nudges tailored to students’ motivation and perceived ability levels”, *British Journal of Educational Technology* (2022).
- [284] Polat, M., “Analysis of multiple-choice versus open-ended questions in language tests according to different cognitive domain levels”, *Novitas-ROYAL (Research on Youth and Language)* **14**, 2, 76–96 (2020).
- [285] PowerSchool, “Schoolology learning”, URL <https://www.powerschool.com/solutions/unified-classroom/schoolology-learning/> (2009).
- [286] Price, T., R. Zhi and T. Barnes, “Evaluation of a data-driven feedback algorithm for open-ended programming.”, International Educational Data Mining Society (2017).
- [287] Prieto, L. P., M. Holenko Dlab, I. Gutiérrez, M. Abdulwahed and W. Balid, “Orchestrating technology enhanced learning: a literature review and a conceptual framework”, *International Journal of Technology Enhanced Learning* **3**, 6, 583–598 (2011).

- [288] Prihar, E., T. Patikorn, A. Botelho, A. Sales and N. Heffernan, “Toward personalizing students’ education with crowdsourced tutoring”, in “Proceedings of the Eighth ACM Conference on Learning@Scale”, pp. 37–45 (2021).
- [289] Prihar, E., T. Patikorn, A. Botelho, A. Sales and N. Heffernan, “Toward personalizing students’ education with crowdsourced tutoring”, in “Proceedings of the Eighth ACM Conference on Learning@Scale”, pp. 37–45 (2021).
- [290] Prihar, E., M. Syed, K. Ostrow, S. Shaw, A. Sales and N. Heffernan, “Exploring common trends in online educational experiments”, in “Proceedings of the 15th International Conference on Educational Data Mining”, p. 27 (2022).
- [291] Prihar, E., M. Syed, K. Ostrow, S. Shaw, A. Sales and N. Heffernan, “Exploring common trends in online educational experiments”, in “Proceedings of the 15th International Conference on Educational Data Mining”, p. 27 (2022).
- [292] Prihar, E., M. Syed, K. Ostrow, S. Shaw, A. Sales and N. Heffernan, “Exploring common trends in online educational experiments”, in “Proceedings of the 15th International Conference on Educational Data Mining”, p. 27–38 (International Educational Data Mining Society, Durham, United Kingdom, 2022).
- [293] Protivínský, T. and D. Münich, “Gender bias in teachers’ grading: What is in the grade”, *Studies in Educational Evaluation* **59**, 141–149 (2018).
- [294] Qiu, Y., Y. Qi, H. Lu, Z. A. Pardos and N. T. Heffernan, “Does time matter? modeling the effect of time with bayesian knowledge tracing.”, in “EDM”, pp. 139–148 (2011).
- [295] Quinlan, K. M., “How emotion matters in four key relationships in teaching and learning in higher education”, *College Teaching* **64**, 3, 101–111 (2016).
- [296] Quinn, D. M., “Experimental evidence on teachers’ racial bias in student evaluation: The role of grading scales”, *Educational Evaluation and Policy Analysis* **42**, 3, 375–392 (2020).
- [297] Quintana, R., C. Quintana, C. Madeira and J. D. Slotta, “Keeping watch: Exploring wearable technology designs for k-12 teachers”, in “Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems”, pp. 2272–2278 (2016).
- [298] Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision”, in “International Conference on Machine Learning”, pp. 8748–8763 (PMLR, 2021).
- [299] Rafferty, A. N., E. Brunskill, T. L. Griffiths and P. Shafto, “Faster teaching via pomdp planning”, *Cognitive science* **40**, 6, 1290–1332 (2016).
- [300] Rauch, D. P. and J. Hartig, “Multiple-choice versus open-ended response formats of reading test items: A two-dimensional irt analysis.”, *Psychological Test and Assessment Modeling* **52**, 354–379 (2010).
- [301] Reich, J. (Harvard University Press, Cambridge, MA, 2020).
- [302] Reimers, N. and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks”, arXiv preprint arXiv:1908.10084 (2019).
- [303] Reio, T., “Prior knowledge, self-directed learning readiness, and curiosity: Antecedents to classroom learning performance”, *International Journal of Self-directed learning* **1**, 1, 18–25 (2004).
- [304] Reio, T. and W. Davis, “Age and gender differences in self-directed learning readiness: A developmental perspective”, *International Journal of Self-Directed Learning* **2**, 1, 40–49 (2005).
- [305] Rhew, E., J. S. Piro, P. Goolkasian and P. Cosentino, “The effects of a growth mindset on self-efficacy and motivation”, *Cogent Education* **5**, 1, 1492337 (2018).

- [306] Riordan, B., A. Horbach, A. Cahill, T. Zesch and C. Lee, “Investigating neural architectures for short answer scoring”, in “Proceedings of the 12th workshop on innovative use of NLP for building educational applications”, pp. 159–168 (2017).
- [307] Ritter, S., J. R. Anderson, K. R. Koedinger and A. Corbett, “Cognitive tutor: Applied research in mathematics education”, *Psychonomic bulletin & review* **14**, 2, 249–255 (2007).
- [308] Rivera-Bergollo, R., S. Baral, A. Botelho and N. Heffernan, “Leveraging auxiliary data from similar problems to improve automatic open response scoring”, *Proceedings of the 15th International Conference on Educational Data Mining* pp. 679–683 (2022).
- [309] Roen, D., “Gender and teacher response to student writing”, in “Gender issues in the teaching of English”, pp. 126–141 (Heinemann, 1992).
- [310] Rogers, C. M., M. D. Smith and J. M. Coleman, “Social comparison in the classroom: the relationship between academic achievement and self-concept.”, *Journal of educational psychology* **70**, 1, 50 (1978).
- [311] Rogers, C. M., M. D. Smith and J. M. Coleman, “Social comparison in the classroom: the relationship between academic achievement and self-concept.”, *Journal of educational psychology* **70**, 1, 50 (1978).
- [312] Roschelle, J., Y. Dimitriadis and U. Hoppe, “Classroom orchestration: synthesis”, *Computers & Education* **69**, 523–526 (2013).
- [313] Roschelle, J., M. Feng, R. F. Murphy and C. A. Mason, “Online mathematics homework increases student achievement”, *AERA open* **2**, 4, 2332858416673968 (2016).
- [314] Roscoe, R. D., L. K. Allen and D. S. McNamara, “Contrasting writing practice formats in a writing strategy tutoring system”, *Journal of Educational Computing Research* **57**, 3, 723–754 (2019).
- [315] Rosen, J. A., E. J. Glennie, B. W. Dalton, J. M. Lennon and R. N. Bozick, *Noncognitive skills in the classroom: New perspectives on educational research* (RTI Press, 2010).
- [316] Rovai, A. P., “Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks”, *Internet and Higher Education* **5**, 4, 319–332 (2002).
- [317] Ruit, K. and P. Carr, “Comparison of student performance on “selected-response” versus “constructed-response” question formats in a medical neuroscience laboratory practical examination”, *The FASEB Journal* **25**, S1, 182–186, URL https://onlinelibrary.wiley.com/doi/10.1096/fasebj.25.1_supplement.182.6 (2011).
- [318] Rummel, A. and R. Feinberg, “Cognitive evaluation theory: A meta-analytic review of the literature”, *Social Behavior and Personality: an international journal* **16**, 2, 147–164 (1988).
- [319] Rushton, S. J., “Teaching and learning mathematics through error analysis”, *Fields Mathematics Education Journal* **3**, 1, 1–12 (2018).
- [320] Sabot, R. and J. Wakeman-Linn, “Grade inflation and course choice”, *Journal of Economic Perspectives* **5**, 1, 159–170 (1991).
- [321] Sajjadi, M. S., M. Alamgir and U. von Luxburg, “Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines”, in “Proceedings of the Third (2016) ACM Conference on Learning @ Scale”, L@S ’16, p. 369–378 (Association for Computing Machinery, New York, NY, USA, 2016), URL <https://doi.org/10.1145/2876034.2876036>.
- [322] Sales, A. C. and J. F. Pane, “Student log-data from a randomized evaluation of educational technology: A causal case study”, *arXiv preprint arXiv:1808.02528* (2018).
- [323] Sams, A. and J. Bergmann, *Flip Your Classroom: Reach Every Student in Every Class Every Day* (2012).
- [324] Savvides, H. and C. Bond, “How does growth mindset inform interventions in primary schools? a systematic literature review”, *Educational Psychology in Practice* **37**, 2, 134–149 (2021).

- [325] Schmidt, F. T., A. Kaiser and J. Retelsdorf, “Halo effects in grading: an experimental approach”, *Educational Psychology* **43**, 2-3, 246–262 (2023).
- [326] Schnepfer, L. C. and L. P. McCoy, “Analysis of misconceptions in high school mathematics”, *Networks: An Online Journal for Teacher Research* **15**, 1, 625–625 (2013).
- [327] Schoenfeld, A. H., “Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (reprint)”, *Journal of education* **196**, 2, 1–38 (2016).
- [328] Segedy, J. R., G. Biswas, E. F. Blackstock and A. Jenkins, “Guided skill practice as an adaptive scaffolding strategy in open-ended learning environments”, in “International Conference on Artificial Intelligence in Education”, pp. 532–541 (Springer, 2013).
- [329] Selent, D. and N. Heffernan, “Reducing student hint use by creating buggy messages from machine learned incorrect processes”, in “International conference on intelligent tutoring systems”, pp. 674–675 (Springer, 2014).
- [330] Seyahi, L. S., S. G. Ozcan, N. Sut, A. Mayer and B. C. Poyraz, “Social and psychiatric effects of covid-19 pandemic and distance learning on high school students: A cross-sectional web-based survey comparing turkey and denmark”, *medRxiv* (2020).
- [331] Shaikh, E., I. Mohiuddin, A. Manzoor, G. Latif and N. Mohammad, “Automated grading for handwritten answer sheets using convolutional neural networks”, in “2019 2nd International conference on new trends in computing sciences (ICTCS)”, pp. 1–6 (Ieee, 2019).
- [332] Shen, J. T., M. Yamashita, E. Prihar, N. Heffernan, X. Wu, B. Graff and D. Lee, “Mathbert: A pre-trained language model for general nlp tasks in mathematics education”, *arXiv preprint arXiv:2106.07340* (2021).
- [333] Shih, B., K. R. Koedinger and R. Scheines, “A response time model for bottom-out hints as worked examples”, *Handbook of educational data mining* pp. 201–212 (2011).
- [334] Siegler, R., “Implications of cognitive science research for mathematics education”, *Colección Digital Eudoxus*, 8 (2009).
- [335] Siegler, R. S., “Strategy choices in addition and subtraction: How do children know what to do?”, *Origins of cognitive skills* (1984).
- [336] Siegler, R. S., “Individual differences in strategy choices: Good students, not-so-good students, and perfectionists”, *Child development* pp. 833–851 (1988).
- [337] Siegler, R. S., *Emerging minds: The process of change in children’s thinking* (Oxford University Press, 1998).
- [338] Silver, E. A., “The nature and use of open problems in mathematics education: Mathematical and pedagogical perspectives.”, *Zentralblatt fur Didaktik der Mathematik/International Reviews on Mathematical Education* **27**, 2, 67–72 (1995).
- [339] Singh, A., S. Karayev, K. Gutowski and P. Abbeel, “Gradescope: A fast, flexible, and fair system for scalable assessment of handwritten work”, in “Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale”, L@S ’17, p. 81–88 (Association for Computing Machinery, New York, NY, USA, 2017), URL <https://doi.org/10.1145/3051457.3051466>.
- [340] Sisk, V. F., A. P. Burgoyne, J. Sun, J. L. Butler and B. N. Macnamara, “To what extent and under which circumstances are growth mind-sets important to academic achievement? two meta-analyses”, *Psychological Science* **29**, 4, 549–571, URL <https://doi.org/10.1177/0956797617739704>, PMID: 29505339 (2018).
- [341] Sison, R. and M. Shimura, “Student modeling and machine learning”, *International Journal of Artificial Intelligence in Education (IJAIED)* **9**, 128–158 (1998).

- [342] Skaalvik, E. M. and S. Skaalvik, "Internal and external frames of reference for academic self-concept", *Educational Psychologist* **37**, 4, 233–244 (2002).
- [343] Skiba, R. J., A. Casey and B. A. Center, "Nonaversive procedures in the treatment of classroom behavior problems", *The Journal of Special Education* **19**, 4, 459–481 (1985).
- [344] Slavin, R. E., "Mastery learning reconsidered", *Review of Educational Research* **57**, 2, 175–213 (1987).
- [345] Smith, M. A. and J. D. Karpicke, "Retrieval practice with short-answer, multiple-choice, and hybrid tests", *Memory* **22**, 7, 784–802 (2014).
- [346] Smith, R., "An overview of the tesseract ocr engine", in "Ninth international conference on document analysis and recognition (ICDAR 2007)", vol. 2, pp. 629–633 (IEEE, 2007).
- [347] Spear, M. G., "The biasing influence of pupil sex in a science marking exercise", *Research in Science & Technological Education* **2**, 1, 55–60 (1984).
- [348] Srihari, S., J. Collins, R. Srihari, P. Babu and H. Srinivasan, "Automated scoring of handwritten essays based on latent semantic analysis", in "International Workshop on Document Analysis Systems", pp. 71–83 (Springer, 2006).
- [349] Stankov, L. and J. Lee, "Quest for the best non-cognitive predictor of academic achievement", *Educational Psychology* **34**, 1, 1–8, URL <https://doi.org/10.1080/01443410.2013.858908> (2014).
- [350] Sugrue, B., N. Webb and J. Schlackman, "The interchangeability of assessment methods in science. cse technical report 474.", Tech. rep., Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA (1998).
- [351] Talib, A., A. M. Bettayeb and R. I. Omer, "Analytical study on the impact of technology in higher education during the age of covid-19: Systematic literature review", *Educ Inf Technol* **26**, 6, 6719–6746 (2021).
- [352] Tenenbaum, G. and E. Goldring, "A meta-analysis of the effect of enhanced instruction: Cues, participation, reinforcement and feedback and correctives on motor skill learning.", *Journal of Research & Development in Education* (1989).
- [353] Thaler, R. H. and C. R. Sunstein, *Nudge* (Yale University Press, 2021).
- [354] Thomas, A. E., "Gender differences in students' physical science motivation: Are teachers' implicit cognitions another piece of the puzzle?", *American Educational Research Journal* **54**, 1, 35–58 (2017).
- [355] Throndsen, I., "Self-regulated learning of basic arithmetic skills: A longitudinal study", *British Journal of Educational Psychology* **81**, 4, 558–578 (2011).
- [356] Tissenbaum, M., C. Matuk, M. Berland, L. Lyons, F. Cocco, M. Linn, J. L. Plass, N. Hajny, A. Olsen, B. Schwendimann *et al.*, "Real-time visualization of student activities to support classroom orchestration", (Singapore: International Society of the Learning Sciences, 2016).
- [357] Topping, K. J., "Trends in peer learning", *Educational Psychology* **25**, 6, 631–645 (2005).
- [358] Vaessen, B. E., F. J. Prins and J. Jeuring, "University students' achievement goals and help-seeking strategies in an intelligent tutoring system", *Computers & Education* **72**, 196–208 (2014).
- [359] VanLehn, K., "Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills.", *The Journal of Mathematical Behavior* (1982).
- [360] VanLehn, K., H. Burkhardt, S. Cheema, S. Kang, D. Pead, A. Schoenfeld and J. Wetzel, "Can an orchestration system increase collaborative, productive struggle in teaching-by-eliciting classrooms?", *Interactive Learning Environments* pp. 1–19 (2019).

- [361] VanLehn, K., S. Cheema, J. Wetzel and D. Pead, “Some less obvious features of classroom orchestration systems”, *Educational technologies: Challenges, applications and learning Outcomes* pp. 73–94 (2016).
- [362] VanLehn, K., S. Siler, C. Murray, T. Yamauchi and W. B. Baggett, “Why do only some events cause learning during human tutoring?”, *Cognition and Instruction* **21**, 3, 209–249 (2003).
- [363] Veenman, M. V., V. Hout-Wolters, H. Bernadette and P. Afflerbach, “Metacognition and learning: Conceptual and methodological considerations”, *Metacognition and learning* **1**, 1, 3–14 (2006).
- [364] Veenman, M. V. J., “Alternative assessment of strategy use with self-report instruments: a discussion”, *Metacognition and Learning* **6**, 2, 205–211 (2011).
- [365] Verbert, K., S. Govaerts, E. Duval, J. L. Santos, F. Van Assche, G. Parra and J. Klerkx, “Learning dashboards: an overview and future research opportunities”, *Personal and Ubiquitous Computing* **18**, 6, 1499–1514 (2014).
- [366] Verweij, D., S. Bakker and B. Eggen, “Fireflies2: Interactive tangible pixels to enable distributed cognition in classroom technologies”, in “Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces”, pp. 260–269 (2017).
- [367] Voltmer, K. and M. von Salisch, “Three meta-analyses of children’s emotion knowledge and their school success”, *Learning and Individual Differences* **59**, 107–118 (2017).
- [368] Vrugt, A. and F. J. Oort, “Metacognition, achievement goals, study strategies and academic achievement: pathways to achievement”, *Metacognition and learning* **3**, 2, 123–146 (2008).
- [369] Walton, D. N., *Plausible argument in everyday conversation* (SUNY Press, 1992).
- [370] Wang, H. and N. C. Hall, “A systematic review of teachers’ causal attributions: Prevalence, correlates, and consequences”, *Frontiers in psychology* **9**, 2305 (2018).
- [371] Wang, X., C. Rose and K. Koedinger, “Seeing beyond expert blind spots: Online learning design for scale and quality”, in “Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems”, p. 1–14 (ACM, Yokohama Japan, 2021), URL <https://dl.acm.org/doi/10.1145/3411764.3445045>.
- [372] Wang, X., S. T. Talluri, C. Rose and K. Koedinger, “Upgrade: Sourcing student open-ended solutions to create scalable learning opportunities”, in “Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale”, L@S ’19 (Association for Computing Machinery, New York, NY, USA, 2019), URL <https://doi.org/10.1145/3330430.3333614>.
- [373] Wang, X., S. T. Talluri, C. Rose and K. Koedinger, “Upgrade: Sourcing student open-ended solutions to create scalable learning opportunities”, in “Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale”, pp. 1–10 (2019).
- [374] Wang, Y., K. Ostrow and N. Heffernan, “Partial credit revisited: Enhancing the efficiency and reliability of group differentiation at scale”, *Student Modeling from Different Aspects* **7**, 6,314, 22 (2016).
- [375] Warikoo, N., S. Sinclair, J. Fei and D. Jacoby-Senghor, “Examining racial bias in education: A new approach”, *Educational Researcher* **45**, 9, 508–514 (2016).
- [376] Weaver, M. R., “Do students value feedback? student perceptions of tutors’ written responses”, *Assessment & Evaluation in Higher Education* **31**, 3, 379–394 (2006).
- [377] Weir, S., J. Kim, K. Z. Gajos and R. C. Miller, “Learnersourcing subgoal labels for how-to videos”, in “Proceedings of the 18th ACM conference on computer supported cooperative work & social computing”, pp. 405–416 (2015).
- [378] Weiser, M. and J. S. Brown, *The coming age of calm technology* (1997).
- [379] Wen, S.-S., “Racial halo on evaluative rating: General or differential?”, *Contemporary Educational Psychology* **4**, 1, 15–19 (1979).

- [380] Whelan, R., “Effective analysis of reaction time data”, *The Psychological Record* **58**, 3, 475–482 (2008).
- [381] Whitehill, J. and M. Seltzer, “A crowdsourcing approach to collecting tutorial videos—toward personalized learning-at-scale”, in “Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale”, pp. 157–160 (2017).
- [382] Wiggins, J. B., K. E. Boyer, A. Baikadi, A. Ezen-Can, J. F. Grafsgaard, E. Y. Ha, J. C. Lester, C. M. Mitchell and E. N. Wiebe, “Javatutor: an intelligent tutoring system that adapts to cognitive and affective states during computer programming”, in “Proceedings of the 46th acm technical symposium on computer science education”, pp. 599–599 (2015).
- [383] Wikström, C. and M. Wikström, “Grade inflation and school competition: an empirical analysis based on the swedish upper secondary schools”, *Economics of education Review* **24**, 3, 309–322 (2005).
- [384] Williams, J. J., J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki and N. Heffernan, “Axis: Generating explanations at scale with learnersourcing and machine learning”, in “Proceedings of the Third (2016) ACM Conference on Learning@ Scale”, pp. 379–388 (2016).
- [385] Winkler, R., S. Hobert, A. Salovaara, M. Söllner and J. M. Leimeister, “Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent”, in “Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems”, pp. 1–14 (2020).
- [386] Wolfe, J. M. and W. Gray, “Guided search 4.0”, *Integrated models of cognitive systems* pp. 99–119 (2007).
- [387] Woodward, J. and L. Howard, “The misconceptions of youth: Errors and their mathematical meaning”, *Exceptional Children* **61**, 2, 126 (1994).
- [388] Wormeli, R., *Fair isn’t always equal: Assessing & grading in the differentiated classroom* (Stenhouse Publishers, 2018).
- [389] Xu, Z., K. Wijekumar, G. Ramirez, X. Hu and R. Irey, “The effectiveness of intelligent tutoring systems on k-12 students’ reading comprehension: A meta-analysis”, *British Journal of Educational Technology* **50**, 6, 3119–3137 (2019).
- [390] Yeager, D. S. and C. S. Dweck, “Mindsets that promote resilience: When students believe that personal characteristics can be developed”, *Educational psychologist* **47**, 4, 302–314 (2012).
- [391] Yeager, D. S. and C. S. Dweck, “What can be learned from growth mindset controversies?”, *American psychologist* **75**, 9, 1269 (2020).
- [392] Yeager, D. S., P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller, E. Tipton, B. Schneider, C. S. Hulleman, C. P. Hinojosa *et al.*, “A national experiment reveals where a growth mindset improves achievement”, *Nature* **573**, 7774, 364–369 (2019).
- [393] Yeager, D. S. and G. M. Walton, “Social-psychological interventions in education: They’re not magic”, *Review of educational Research* **81**, 2, 267–301 (2011).
- [394] Young, R. M. and T. O’Shea, “Errors in children’s subtraction”, *Cognitive Science* **5**, 2, 153–177 (1981).
- [395] Zhang, L., Y. Huang, X. Yang, S. Yu and F. Zhuang, “An automatic short-answer grading model for semi-open-ended questions”, *Interactive learning environments* **30**, 1, 177–190 (2022).
- [396] Zhang, M., S. Baral, N. Heffernan and A. Lan, “Automatic short math answer grading via in-context meta-learning”, arXiv preprint arXiv:2205.15219 (2022).
- [397] Zhao, S., Y. Zhang, X. Xiong, A. Botelho and N. Heffernan, “A memory-augmented neural model for automated grading”, in “Proceedings of the fourth (2017) ACM conference on learning@ scale”, pp. 189–192 (2017).
- [398] Zhu, M., O. L. Liu and H.-S. Lee, “The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing”, *Computers & Education* **143**, 103668 (2020).

Appendices

Chapter A

SUPPLEMENTARY MATERIALS FOR CHAPTER 8 “EXPLORING THE INFLUENCE OF ANONYMITY AND PRIOR-PERFORMANCE ON TEACHER GRADING BEHAVIOR”

A.1 Study Design

This section elaborates on the visualization used per condition in the 2×2 experimental design. As the two factors in the experiment were student identity and prior performance information. There are four cells in the experiment.

The first cell where students are anonymized and prior performance is not provided as illustrated in figure A.1. The teacher have access to the students response. They can examine the problem and prior subparts of the problem if it is a multipart problem when assessing the response. The teachers are required to provide a grade and a feedback.

The screenshot displays a grading interface. On the left, under the heading "Problem Body", there are three parts of a problem. "Part: 1" is a grey header. "Part: 2" is a blue header with the text "Write a ratio for 4 batches of this recipe." and "Type your answer in the format: $\$4$ with no spaces." "Part: 3 (Current problem)" is a blue header with the text "Explain why we can say that any two of these three ratios are equivalent." On the right, the student's response is shown in a text box: "They are equivalent because they are both even". Below the response, there are fields for "Score:" (a dropdown menu) and "Feedback:" (a text input field). A blue button labeled "Score Next Response" is at the bottom.

Figure A.1: A screenshot from the grading tool where a teacher is grading a response where the student is anonymized and prior performance information is not provided. We display the problem body, prior subparts, and student's response.

The second cell where student names (pseudonyms) are provided but their prior performance is not provided as illustrated in figure A.2. The teacher have access to the students response and are required to provide a grade and a feedback.

The second cell where student are anonymized but their prior performance information is not provided as illustrated in figure A.3. The average correctness of the students in the past 5 assignments prior to working on the current assignment is used as a proxy for prior performance information indicator. The teacher have access to the students response and are required to provide a grade and a feedback.

The third cell where student names (pseudonyms) and their prior performance information is not provided as illustrated in figure A.4. The average correctness of the students in the past 5 assignments prior to working on the current assignment is used as a proxy for prior performance information indicator. The teacher have access to the students response and are required to provide a grade and a feedback.

A.2 Additional Materials for Replication

We compute intra-rater reliability for category 1 teachers by comparing the anonymized scores for responses with the original scores prior to the experiment. Among all the teachers, Teacher 3 had significantly lower levels of intra-rater reliability and relaxed intra-rater reliability scores compared to their peers. These findings indicate that the teachers disagree with themselves when student responses are anonymized.

When the comparison was extended to include all 75 problems the intra-rater reliability for teacher 3 remained low. Further exploration of teacher 3 data had some indication of personalization where the teacher feedback, when provided, was indicative of *considerate grading behavior*. There were several instances where the teacher only provided grades without any feedback when they originally graded the responses. Some examples of such feedback is provided in table A.2.

Problem Body

Part: 1

Andre and Noah started tracking their savings at the same time. Andre started with \$15 and deposits \$5 per week. Noah started with \$2.50 and deposits \$7.50 per week. The graph of Noah's savings is given and his equation is $y = 7.5x + 2.5$, where x represents the number of weeks and y represents his savings.

Write the equation for Andre's savings and graph it alongside Noah's. What does the intersection point mean in this situation?

Write the equation for Andre's savings.
 $y =$ _____
 Use x as your variable.

Copied for free from openupresources.org

Part: 2

Part: 3 (Current problem)

What does the intersection point of Noah's and Andre's savings mean in this situation?

Student Name: Brianna Booker
 Brianna's Response:

Score: Feedback:

[Score Next Response](#)

Figure A.2: A screenshot from the grading tool where a teacher is grading a response where the student identity (pseudonym) is provided but prior performance information is hidden. We display the problem body, prior sub-parts, and student's response.

Problem Body

Part: 1 (Current problem)

You may use a scientific calculator, but not a graphing calculator.

A community theater uses the function $p(d) = -4d^2 + 200d - 100$ to model the profit (in dollars) expected in a weekend when the tickets to a comedy show are priced at d dollars each.

Write and solve an equation to find out the prices at which the theater would earn \$1500 in profit from the comedy show each weekend.

First, write the equation.

Write the equation using the "WIRIS editor" button

Student Name: Anonymized
 Student performance in the last 5 assignments:
 Current Assignment is Assignment 0:

Students's Response:

Score: Feedback:

[Score Next Response](#)

Figure A.3: A screenshot from the grading tool where a teacher is grading a response where students are anonymized but we provide their prior performance. We display the problem body, prior sub-parts, student's response and average performance on prior 5 assignments.

Figure A.5 displays a screenshot capturing the Google search results for the term “Equivalent Ratios.” As illustrated in Example 1 from Table A.2, it is noteworthy that the teacher’s initial feedback was *Please do not copy and paste responses from a Google Search*. This feedback implies that the teacher possessed a nuanced understanding of the student’s abilities and inclinations. Being cognizant of the tendencies among some students, the teacher was vigilant in assessing student responses, actively monitored submissions for such behavior, and provided targeted feedback and grading to deter such behavior in future assignments.

Problem Body

Part: 1

A recipe for orange water says, "Mix 3 teaspoons yellow water with 1 teaspoon red water." For this recipe, we might say, "The ratio of teaspoons of yellow water to teaspoons of red water is 3:1."

Write a ratio for 2 batches of this recipe.

Type your answer in the format: **3:4** with no spaces.

Part: 2

Part: 3 (Current problem)

Explain why we can say that any two of these three ratios are equivalent.

Student Name: Jaylen Alston

Student performance in the last 5 assignments:
Current Assignment is Assignment 0:

Jaylen's Response:

They mutiply by 3 and 2

Score: Feedback:

[Score Next Response](#)

Figure A.4: A screenshot from the grading tool where a teacher is grading a response where the student identity (pseudonym) and prior performance information is provided. We display the problem body, prior sub-parts, student's response and average performance on prior 5 assignments.

Table A.1: Comparison of Original and Anonymized Scores for intra-rater Reliability of Category 1 Teachers' Response Grades: Replicating Findings from Prior Work [144] on All Responses

Teacher	Responses N	intra-rater Reliability	
		Original vs Anon. Score	Relaxed intra-rater Reliability Original vs Anon. Score
Teacher 1	75	0.66	0.96
Teacher 2	75	0.66	0.94
Teacher 3	75	0.33	0.61
Teacher 4	75	0.56	0.76
Teacher 5	75	0.61	0.77
Teacher 6	75	0.68	0.78
Teacher 7	75	0.51	0.75

Table A.2: Examining some of the grades and feedback of teacher 3 with relatively low intra-rater reliability.
Example 1:

Answer:	Equivalent ratios are ratios that make the same comparison of numbers. Two ratios are equivalent if one can be expressed as a multiple of the other. ... In this example, that ratio is 1 : 2 : 4.	
Condition	Score	Feedback
Original	1	Please do not copy and paste responses from a Google search
Anonymized	4	Nice response

Example 2:

Answer:	i got 12:4 because in order to write a ratio for 4 batches, you must multiply both numbers in the original ratio by 4	
Condition	Score	Feedback
Original	0	left blank
Anonymized	3	This is true, but does this apply to any of the ratios in the problem?

Example 3:

Answer:	por que el numero nunca cambia	
Condition	Score	Feedback
Original	2	N/A
Anonymized	2	Creo que entiendes las proporciones equivalentes, pero tu explicación necesita un poco más.

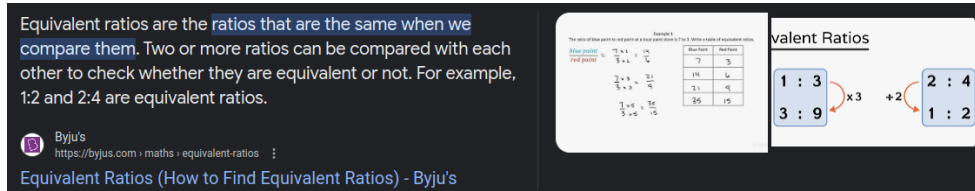


Figure A.5: A screenshot of the Google search results for the term “Equivalent Ratios.”

A.2.1 Intra-rater reliability across condition per teacher

The reference to the cells is visually presented in the experimental visualization of the factorial experimental setup in figure A.6. Codes for comparing across cells in a 2×2 factorial design:

- anon wo prior v. not anon wo prior : (00, 10)
- anon wo prior v. anon w prior : (00, 01)
- anon wo prior v. not anon w prior : (00, 11)
- anon w prior v. not anon wo prior : (01, 10)
- anon w prior v. not anon w prior : (01, 11)
- not anon wo prior v. not anon w prior : (10, 11)

		Prior Performance Information	
		Without Prior Info (0)	With Prior Info (1)
Anonymization	Anonymized (0)	Subsample Category 1 N >= 15 (Cell 00)	Subsample Category 3 N >= 15 (Cell 01)
	Not Anonymized (1)	Subsample Category 2 N >= 15 (Cell 10)	Subsample Category 4 N >= 15 (Cell 11)

Figure A.6: A visual representation of the 2×2 factorial randomized control trial.

We also explored the intra-rater reliability in teacher grading behavior to analyze the consistency in teacher grading behaviors across conditions in the randomized control trial. As there are 4 cells in the factorial RCT this results in the possibility of making ${}^4C_2 = 6$ comparison when evaluating teacher grades across conditions. These comparisons using regular Cohen’s Kappa and relaxed Cohen’s Kappa (score off by 1) scores are presented in table A.3. We observed a relatively high agreement scores among teachers with only two teachers T3, T16 exhibiting low agreement across the different conditions, however when the constraints were relaxed they had a strong intra-rater reliability as well, indicating that their grading practices while more variant than their peers is often only off by a single point across conditions.

Table A.3: Regular and Relaxed (Off by 1) intra-rater reliability across condition per teacher. The values ≤ 0.5 intra-rater reliability have been marked in red.

teacher xid	category	type	(00, 10)	(00, 01)	(00, 11)	(01,10)	(01,11)	(10,11)
T1	0	regular	0.7	0.74	0.71	0.69	0.73	0.75
		relaxed	0.76	0.82	0.72	0.75	0.77	0.77
T2	0	regular	0.7	0.76	0.64	0.67	0.62	0.64
		relaxed	0.78	0.84	0.77	0.73	0.83	0.75
T3	0	regular	0.47	0.51	0.53	0.57	0.5	0.51
		relaxed	0.72	0.69	0.66	0.75	0.6	0.66
T4	0	regular	0.63	0.61	0.7	0.67	0.62	0.64
		relaxed	0.67	0.74	0.74	0.79	0.74	0.66
T5	0	regular	0.69	0.67	0.65	0.68	0.67	0.65
		relaxed	0.76	0.79	0.77	0.77	0.82	0.77
T6	0	regular	0.73	0.72	0.73	0.74	0.76	0.69
		relaxed	0.77	0.74	0.75	0.81	0.8	0.77
T7	0	regular	0.7	0.64	0.7	0.62	0.6	0.62
		relaxed	0.77	0.8	0.79	0.81	0.79	0.76
T8	1	regular	0.73	0.72	0.77	0.74	0.76	0.73
		relaxed	0.75	0.72	0.77	0.76	0.78	0.77
T9	1	regular	0.76	0.77	0.75	0.72	0.82	0.76
		relaxed	0.83	0.83	0.82	0.8	0.86	0.83
T10	1	regular	0.64	0.6	0.61	0.59	0.56	0.58
		relaxed	0.77	0.75	0.79	0.81	0.75	0.78
T11	1	regular	0.82	0.79	0.8	0.76	0.85	0.8
		relaxed	0.82	0.84	0.84	0.77	0.87	0.85
T12	1	regular	0.78	0.8	0.77	0.81	0.8	0.83
		relaxed	0.79	0.85	0.86	0.87	0.81	0.9
T13	1	regular	0.85	0.86	0.83	0.88	0.83	0.85
		relaxed	0.89	0.92	0.92	0.96	0.94	0.92
T14	1	regular	0.73	0.73	0.69	0.75	0.74	0.8
		relaxed	0.86	0.78	0.82	0.82	0.84	0.86
T15	2	regular	0.79	0.79	0.79	0.76	0.79	0.85
		relaxed	0.91	0.85	0.89	0.91	0.93	0.91
T16	2	regular	0.34	0.39	0.26	0.35	0.4	0.31
		relaxed	0.72	0.74	0.73	0.74	0.75	0.76
T17	2	regular	0.74	0.76	0.75	0.73	0.78	0.7
		relaxed	0.83	0.86	0.82	0.8	0.83	0.86
T18	2	regular	0.92	0.89	0.9	0.87	0.83	0.91
		relaxed	0.94	0.93	0.96	0.91	0.92	0.94

A.3 Main Effects

The computation of main effects in a 2×2 factorial design can be simplified by utilizing a single model instead of employing two separate models. Effect coding, using (-0.5, 0.5) codes for the factors, presents an alternative to the conventional dummy coding method. In this section, we elaborate on the computation of main effects using effect coding. Furthermore, we compare the results obtained through effect coding with those obtained using the conventional approach of dummy coding, which was adopted in this study. It is noteworthy that the decision to employ dummy coding was driven by the author's aim to facilitate a more intuitive interpretation of the results, given the prevalence of dummy coding as the commonly used methodological approach.

Main Equation:

Here, X_1 , and X_2 are the two factors:

$$G = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Main Effects using Effect Coding:

Using (-0.5, 0.5) for effect coding to estimate the main effects,

if $X_1 = -0.5$:

$$E[G|X_1 = -0.5] = \alpha_1 + \beta_1(-0.5) + \beta_2 \frac{0.5-0.5}{2} + \beta_3(-0.5) \frac{0.5-0.5}{2}$$

$$E[G|X_1 = -0.5] = \alpha_1 - 0.5\beta_1$$

if $X_1 = 0.5$:

$$E[G|X_1 = 0.5] = \alpha_1 + 0.5\beta_1$$

Now, computing the difference:

$$E[G|X_1 = 0.5] - E[G|X_1 = -0.5] = \beta_1$$

i.e., the main effect of factor X_1 is β_1

Similarly,

$$E[G|X_2 = 0.5] - E[G|X_2 = -0.5] = \beta_2$$

i.e., the main effect of factor X_2 is β_2

The models for estimating the main effects, utilizing both dummy coding and effect coding (-0.5, 0.5) to compare the results across different approaches, are presented in table A.4.

Table A.4: Comparing the estimation of main effects when using dummy coding compared with when using effect coding.

Predictors	Dummy Coded grade			Dummy Coded grade			Dummy Coded Interaction grade			Effect Coding (-0.5, 0.5) grade		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	2.10	1.86 – 2.34	<0.001	2.12	1.88 – 2.36	<0.001	2.11	1.86 – 2.35	<0.001	2.11	1.87 – 2.35	<0.001
experiment anonymized [ethnic_names]	0.02	-0.05 – 0.09	0.528				0.03	-0.07 – 0.13	0.536			
experiment prior performance [with_prior_performance]				-0.02	-0.09 – 0.05	0.597	-0.01	-0.11 – 0.09	0.840			
experiment anonymized [ethnic_names] × experiment prior performance [with_prior_performance]							-0.02	-0.16 – 0.12	0.807			
not anonymized effect coded 05 05										0.02	-0.05 – 0.09	0.528
w prior performance effect coded 05 05										-0.02	-0.09 – 0.05	0.597
not anonymized effect coded 05 05 × w prior performance effect coded 05 05										-0.02	-0.16 – 0.12	0.807
Random Effects												
σ^2	1.79			1.79			1.79			1.79		
τ_{00}	0.35	problem_id		0.35	problem_id		0.35	problem_id		0.35	problem_id	
	0.10	teacher_xid		0.10	teacher_xid		0.10	teacher_xid		0.10	teacher_xid	
ICC	0.20			0.20			0.20			0.20		
N	18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid	
	57	problem_id		57	problem_id		57	problem_id		57	problem_id	
Observations	5400			5400			5400			5400		
Marginal R ² / Conditional R ²	0.000 / 0.201			0.000 / 0.201			0.000 / 0.201			0.000 / 0.201		

From table A.4 the utilization of effect coding facilitates the estimation of main effects for both factors by leveraging the β estimates obtained from the model using effect coding. It is important to acknowledge that while the coefficient from the dummy coded interaction model can be employed to compute the main effects and sub-effects, interpreting the β estimates in the dummy coded interaction model as main effects is a misinterpretation that is often made. The formula to predict the main effects using the dummy coded interaction model is presented in the following section exploring sub-effects.

A.4 Sub Effects

The exploration of the sub-effects is reported in the main report of the paper, specifically in Table 8.8. Here, we will first describe how the sub-effects comparing various conditions can be calculated using the β coefficients from the regression analysis. Subsequently, in the following subsections, we present the regression analyses that were employed to investigate the sub-effects across gender and ethnicity, both individually and in combination. The results from the tables were utilized in generating the plots in figures in section 8.6.

The formula to predict the main effects and sub-effects using the dummy coded interaction model is presented in the following derivation:

For Main Effects with Dummy Coding:

Here, X_1 , and X_2 are the two factors:

$$G = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Using (0,1) for dummy coding, if $X_1 = 0$:

$$E[G|X_1 = 0] = \alpha_1 + \beta_1(0) + \beta_2 \frac{1-0}{2} + \beta_3(0) \frac{1-0}{2}$$

$$E[G|X_1 = 0] = \alpha_1 + 0.5\beta_2$$

if $X_1 = 1$:

$$E[G|X_1 = 1] = \alpha_1 + \beta_1 + 0.5\beta_2 + 0.5\beta_3$$

Now, for main effect of X_1 computing the difference:

$$E[G|X_1 = 1] - E[G|X_1 = 0] = \beta_1 + 0.5\beta_3$$

Similarly, for main effect of X_2

$$E[G|X_2 = 1] - E[G|X_2 = 0] = \beta_2 + 0.5\beta_3$$

For Sub Effects with Dummy Coding:

$$E[G|X_1 = 0, X_2 = 0] = \alpha_1 + \beta_1(0) + \beta_2(0) + \beta_3(0)(0)$$

$$E[G|X_1 = 0, X_2 = 0] = \alpha_1$$

$$E[G|X_1 = 1, X_2 = 0] = \alpha_1 + \beta_1(1) + \beta_2(0) + \beta_3(1)(0)$$

$$E[G|X_1 = 1, X_2 = 0] = \alpha_1 + \beta_1$$

$$E[G|X_1 = 0, X_2 = 1] = \alpha_1 + \beta_1(0) + \beta_2(1) + \beta_3(0)(1)$$

$$E[G|X_1 = 0, X_2 = 1] = \alpha_1 + \beta_2$$

$$E[G|X_1 = 1, X_2 = 1] = \alpha_1 + \beta_1(1) + \beta_2(1) + \beta_3(1)(1)$$

$$E[G|X_1 = 1, X_2 = 1] = \alpha_1 + \beta_1 + \beta_2 + \beta_3$$

Now,

$$\text{sub effect between (1,0) - (0,0) = } \beta_1$$

$$\text{sub effect between (0,1) - (0,0) = } \beta_2$$

$$\text{sub effect between (1,1) - (0,0) = } \beta_1 + \beta_2 + \beta_3$$

$$\text{sub effect between (1,0) - (0,1) = } \beta_1 - \beta_2$$

$$\text{sub effect between (1,0) - (1,1) = } -\beta_2 - \beta_3$$

$$\text{sub effect between (0,1) - (1,1) = } -\beta_1 - \beta_3$$

What is β_3 ?

$$(E[G|X_1 = 0, X_2 = 0] + E[G|X_1 = 1, X_2 = 1]) - (E[G|X_1 = 1, X_2 = 0] + E[G|X_1 = 0, X_2 = 1])$$

$$(\alpha_1 + \alpha_1 + \beta_1 + \beta_2 + \beta_3) - (\alpha_1 + \beta_1 + \alpha_1 + \beta_2)$$

$$\beta_3$$

Hence,

$$\beta_3 = [(0,0)+(1,1)] - [(1,0)+(0,1)]$$

A.4.1 Sub Effects and Learner Gender

We explore heterogeneity in teacher grading behavior due to student identity, prior performance information based on the inferred gender using the pseudonyms separately. In order to do this we first examine the variance in the quality of student responses between boy and girl learners by comparing the scores teachers gave the responses in the baseline condition (anonymized no prior performance info.). While responses assigned to girls did receive lower scores than boys there was no significant difference between the two groups as presented in table A.5.

Table A.5: Comparing sub-effects across genders using anonymized as the baseline to compare the distribution of grades post randomization.

<i>Predictors</i>	grade		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	2.19	1.95 – 2.43	<0.001
[anonymized_no_prior_performance_girl]	-0.09	-0.24 – 0.06	0.231
Random Effects			
σ^2	1.87		
τ_{00} problem_id	0.25		
τ_{00} teacher_xid	0.08		
ICC	0.15		
N teacher_xid	18		
N problem_id	57		
Observations	1350		
Marginal R ² / Conditional R ²	0.001 / 0.152		

We now analyze the influence of student identity and prior performance information on teacher grades when the pseudonym represents a girl compared to when it represents a boy. As presented in table A.6, the findings indicate that there were no significant changes observed in teacher grades attributed to student identity, prior performance information, or their combination.

A.4.2 Sub Effects and Learner Ethnicity

We explore the heterogeneity in teacher grading behavior attributed to student identity and prior performance information, focusing on inferred ethnicity using pseudonyms. To begin, we examine the variance in the quality of student responses among different ethnicities by comparing the scores assigned by teachers in the baseline condition (anonymized, without prior performance information). When using African American students as the reference group, we observed a significant difference in the grades received by Asian, Caucasian, and Middle Eastern students for both genders as presented in table A.7. However, there was no significant difference observed among the South Asian and Hispanic groups. It is important to note that these findings may be a result of chance, as we were unable to control for all possible permutations of response assignments. Furthermore, if we identify differences across the four conditions within any of the ethnic groups, a relative score (increase or decrease) would need to be calculated to estimate the impact of specific learner ethnicity in comparison to other ethnicities.

Table A.6: Comparing sub-effects across conditions for different genders separately.

<i>Predictors</i>	grade			grade		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	2.05	1.75 – 2.35	<0.001	2.13	1.87 – 2.38	<0.001
[anonymized_with_prior_performance_boy]	-0.02	-0.17 – 0.14	0.815			
[ethnic_names_no_prior_performance_boy]	0.02	-0.13 – 0.18	0.796			
[ethnic_names_with_prior_performance_boy]	-0.04	-0.19 – 0.12	0.622			
[anonymized_with_prior_performance_girl]				-0.00	-0.13 – 0.12	0.940
[ethnic_names_no_prior_performance_girl]				0.04	-0.09 – 0.17	0.548
[ethnic_names_with_prior_performance_girl]				0.03	-0.10 – 0.16	0.626
Random Effects						
σ^2	1.68			1.75		
τ_{00}	0.56	problem_id		0.38	problem_id	
	0.11	teacher_xid		0.10	teacher_xid	
ICC	0.28			0.21		
N	18	teacher_xid		18	teacher_xid	
	56	problem_id		57	problem_id	
Observations	2160			3240		
Marginal R ² / Conditional R ²	0.000 / 0.285			0.000 / 0.214		

Table A.7: Comparing sub-effects across ethnicities using anonymized as the baseline to compare the distribution of grades post randomization.

<i>Predictors</i>	<i>Estimates</i>	grade	
		<i>CI</i>	<i>p</i>
(Intercept)	2.35	2.09 – 2.62	< 0.001
[anonymized_no_prior_performance_Asian]	-0.33	-0.59 – -0.07	0.014
[anonymized_no_prior_performance_Caucasian]	-0.34	-0.58 – -0.11	0.005
[anonymized_no_prior_performance_Hispanic]	-0.11	-0.35 – 0.12	0.343
[anonymized_no_prior_performance_MiddleEastern]	-0.39	-0.66 – -0.13	0.003
[anonymized_no_prior_performance_SouthAsian]	-0.23	-0.49 – 0.03	0.081
Random Effects			
σ^2	1.86		
τ_{00} problem_id	0.25		
τ_{00} teacher_xid	0.08		
ICC	0.15		
N teacher_xid	18		
N problem_id	57		
Observations	1350		
Marginal R ² / Conditional R ²	0.009 / 0.158		

In light of the observed differences in the baseline condition across ethnicities, an analysis was conducted to identify the source of this imbalance despite the randomization process. It was found that the cause of the imbalance is primarily attributable to the responses assigned to teachers in Category 0. Notably, the randomly assigned responses given to teachers in Category 0 averaged higher grades compared to other random samples allocated to teachers in Categories 1 and 2. When this factor is combined with the reality that all teachers in Category 0 were assigned the same set of responses, it becomes clear that the unbalanced sample assigned to Category 0 is the primary driver of the imbalance observed in the resulting baseline conditions as presented in table A.8.

Table A.8: Comparing sub-effects across ethnicities using anonymized as the baseline to compare the distribution of grades post randomization.

<i>Predictors</i>	Category 0 grade			Category 1 grade			Category 2 grade		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	2.85	2.38 – 3.32	<0.001	2.18	1.83 – 2.53	<0.001	2.23	1.79 – 2.67	<0.001
[anonymized_no_prior_performance_Asian]	-0.41	-0.77 – -0.04	0.030	-0.16	-0.60 – 0.29	0.488	-0.46	-1.08 – 0.15	0.138
[anonymized_no_prior_performance_Caucasian]	-0.67	-1.00 – -0.34	<0.001	-0.12	-0.51 – 0.28	0.564	-0.19	-0.74 – 0.37	0.509
[anonymized_no_prior_performance_Hispanic]	-0.02	-0.35 – 0.31	0.896	-0.16	-0.55 – 0.24	0.440	-0.19	-0.74 – 0.37	0.508
[anonymized_no_prior_performance_MiddleEastern]	-0.80	-1.17 – -0.44	<0.001	-0.07	-0.51 – 0.37	0.766	-0.29	-0.90 – 0.33	0.358
[anonymized_no_prior_performance_SouthAsian]	-0.56	-0.92 – -0.20	0.003	-0.21	-0.65 – 0.23	0.351	0.29	-0.33 – 0.90	0.360
Random Effects									
σ^2	1.42			2.04			2.26		
τ_{00}	0.18 _{teacher_xid}			0.31 _{problem_id}			0.12 _{problem_id}		
	0.09 _{problem_id}			0.00 _{teacher_xid}			0.02 _{teacher_xid}		
ICC	0.16						0.06		
N	7 _{teacher_xid}			7 _{teacher_xid}			4 _{teacher_xid}		
	5 _{problem_id}			28 _{problem_id}			27 _{problem_id}		
Observations	525			525			300		
Marginal R ² / Conditional R ²	0.056 / 0.205			0.002 / NA			0.019 / 0.077		

We now analyze the influence of student identity and prior performance information on teacher grades when the pseudonym could be used to infer learner ethnicity. As presented in table A.9, we did not observe any significant change due to student identity, prior performance info or both on the grades of the teacher within each ethnic sub-groups.

Table A.9: Comparing sub-effects across conditions for different ethnicities separately.

Predictors	grade			grade			grade			grade					
	Estimates	CI	P	Estimates	CI	P	Estimates	CI	P	Estimates	CI	P			
(Intercept)	2.41	2.04 – 2.78	<0.001	1.96	1.53 – 2.39	<0.001	1.90	1.52 – 2.28	<0.001	2.20	1.82 – 2.57	<0.001	2.11	1.68 – 2.53	<0.001
[anonymized_with_prior_performance_AfricanAmerican]	-0.06	-0.26 – 0.14	0.585												
[ethnic_names_no_prior_performance_AfricanAmerican]	0.03	-0.17 – 0.23	0.799												
[ethnic_names_with_prior_performance_AfricanAmerican]	0.01	-0.18 – 0.21	0.884												
[anonymized_with_prior_performance_Asian]				0.04	-0.21 – 0.28	0.754									
[ethnic_names_no_prior_performance_Asian]				0.09	-0.15 – 0.34	0.447									
[ethnic_names_with_prior_performance_Asian]				0.05	-0.19 – 0.29	0.687									
[anonymized_with_prior_performance_Caucasian]				-0.08	-0.30 – 0.14	0.486									
[ethnic_names_no_prior_performance_Caucasian]				0.00	-0.22 – 0.22	0.974									
[ethnic_names_with_prior_performance_Caucasian]				-0.01	-0.23 – 0.21	0.921									
[anonymized_with_prior_performance_Hispanic]							-0.03	-0.23 – 0.18	0.806						
[ethnic_names_no_prior_performance_Hispanic]							-0.08	-0.29 – 0.13	0.441						
[ethnic_names_with_prior_performance_Hispanic]							-0.03	-0.24 – 0.18	0.779				0.11	-0.13 – 0.34	0.379
[anonymized_with_prior_performance_MiddleEastern]													0.20	-0.04 – 0.44	0.096
[ethnic_names_no_prior_performance_MiddleEastern]													0.10	-0.14 – 0.34	0.405
[ethnic_names_with_prior_performance_MiddleEastern]															
[anonymized_with_prior_performance_MiddleEastern]															
[ethnic_names_with_prior_performance_SouthAsian]													0.02	-0.22 – 0.26	0.891
[anonymized_with_prior_performance_SouthAsian]													0.02	-0.22 – 0.26	0.855
[ethnic_names_no_prior_performance_SouthAsian]													-0.08	-0.32 – 0.16	0.495
[ethnic_names_with_prior_performance_SouthAsian]															
Random Effects															
σ^2	1.40			1.39			1.68			1.51			1.29		
τ_{00}	0.71	problem_id		1.09	problem_id		0.52	problem_id		0.79	problem_id		0.97	problem_id	
	0.20	teacher_xid		0.18	teacher_xid		0.27	teacher_xid		0.18	teacher_xid		0.26	teacher_xid	
ICC	0.39			0.48			0.32			0.39			0.49		
N	18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid	
	50	problem_id		47	problem_id		48	problem_id		47	problem_id		47	problem_id	
Observations	1080			720			1080			1080			720		
Marginal R ² / Conditional R ²	0.000 / 0.392			0.000 / 0.478			0.000 / 0.389			0.000 / 0.487			0.001 / 0.486		

A.4.3 Sub Effects, Learner Gender and Ethnicity

We investigate the heterogeneity in teacher grading behavior attributed to student identity and prior performance information based on the inferred ethnicity using pseudonyms. To accomplish this, we begin by examining the variance in the quality of student responses across different ethnicities for each gender. We achieve this by comparing the scores assigned by teachers in the baseline condition (anonymized, without prior performance information). As presented in table A.10 when African American students are used as the reference group for both genders, we did not observe any significant differences among boys. However, there was a significant difference in the grades received by South Asian, Caucasian, and Middle Eastern girls, while no significant difference was observed between African American, Asian and Hispanic girls. It is important to note that these findings may purely be a function of chance, as we were unable to control for all possible permutations of response assignments. Moreover, if differences are found across the four conditions within any of the groups, it would be necessary to calculate a relative score (increase or decrease) to estimate the impact of a specific learner ethnicity in comparison to other ethnicities.

Table A.10: Comparing sub-effects across ethnicities for both genders using anonymized as the baseline to compare the distribution of grades post randomization.

Predictors	grade			grade		
	Estimates	CI	p	Estimates	CI	p
(Intercept)	2.48	2.20 – 2.76	<0.001	2.13	1.76 – 2.49	<0.001
anon prior genderethnicity[anonymized_no_prior_performance_girl_Asian]	-0.32	-0.68 – 0.03	0.076			
anon prior genderethnicity[anonymized_no_prior_performance_girl_Caucasian]	-0.59	-0.88 – -0.30	<0.001			
anon prior genderethnicity[anonymized_no_prior_performance_girl_Hispanic]	-0.19	-0.48 – 0.10	0.209			
anon prior genderethnicity[anonymized_no_prior_performance_girl_MiddleEastern]	-0.62	-0.98 – -0.27	0.001			
anon prior genderethnicity[anonymized_no_prior_performance_girl_SouthAsian]	-0.37	-0.72 – -0.01	0.043			
anon prior genderethnicity[anonymized_no_prior_performance_boy_Asian]				-0.31	-0.71 – 0.10	0.135
anon prior genderethnicity[anonymized_no_prior_performance_boy_Caucasian]				0.15	-0.26 – 0.56	0.460
anon prior genderethnicity[anonymized_no_prior_performance_boy_Hispanic]				0.01	-0.40 – 0.42	0.958
anon prior genderethnicity[anonymized_no_prior_performance_boy_MiddleEastern]				-0.08	-0.48 – 0.32	0.707
anon prior genderethnicity[anonymized_no_prior_performance_boy_SouthAsian]				-0.13	-0.53 – 0.27	0.519
Random Effects						
σ^2	1.89			1.75		
τ_{00}	0.17	problem_id		0.39	problem_id	
	0.07	teacher_xid		0.06	teacher_xid	
ICC	0.11			0.20		
N	18	teacher_xid		18	teacher_xid	
	57	problem_id		56	problem_id	
Observations	810			540		
Marginal R ² / Conditional R ²	0.025 / 0.135			0.009 / 0.211		

We now analyze the influence of student identity and prior performance information on teacher grades when the pseudonym could be used to infer learner ethnicity and gender. We did not observe any significant change due to student identity, prior performance info or both on the grades of the teacher. Table A.11 presents the difference in grades across condition for boys from different ethnicities where as table A.12 presents the difference in grades across condition for girls from different ethnicities.

Table A.11: Comparing sub-effects across conditions for different ethnicities separately for boys.

Predictors	grade			grade			grade			grade								
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p						
(Intercept)	2.15	1.64–2.67	<0.001	2.01	1.45–2.56	<0.001	2.26	1.77–2.75	<0.001	2.01	1.44–2.58	<0.001	2.05	1.50–2.60	<0.001	2.04	1.56–2.52	<0.001
[anonymized_with_prior_performance_boy_African American]	-0.00	-0.29–0.29	1.000															
[ethnic_names_no_prior_performance_boy_African American]	0.02	-0.26–0.31	0.878															
[ethnic_names_with_prior_performance_boy_African American]	0.03	-0.25–0.32	0.819															
[anonymized_with_prior_performance_boy_Asian]				0.09	-0.22–0.40	0.570												
[ethnic_names_no_prior_performance_boy_Asian]				0.12	-0.19–0.43	0.435												
[ethnic_names_with_prior_performance_boy_Asian]				0.01	-0.30–0.32	0.943												
[anonymized_with_prior_performance_boy_Caucasian]							-0.18	-0.53–0.18	0.324									
[ethnic_names_no_prior_performance_boy_Caucasian]							0.00	-0.35–0.35	1.000									
[ethnic_names_with_prior_performance_boy_Caucasian]							-0.14	-0.50–0.21	0.422									
[anonymized_with_prior_performance_boy_Hispanic]										-0.00	-0.26–0.26	1.000						
[ethnic_names_no_prior_performance_boy_Hispanic]										-0.12	-0.38–0.13	0.350						
[ethnic_names_with_prior_performance_boy_Hispanic]										-0.02	-0.28–0.23	0.865						
[anonymized_with_prior_performance_boy_Middle Eastern]													0.09	-0.18–0.35	0.511			
[ethnic_names_no_prior_performance_boy_Middle Eastern]													0.18	-0.09–0.44	0.189			
[ethnic_names_with_prior_performance_boy_Middle Eastern]													0.10	-0.17–0.37	0.459			
[anonymized_with_prior_performance_boy_South Asian]																-0.11	-0.47–0.25	0.541
[ethnic_names_no_prior_performance_boy_South Asian]																-0.08	-0.43–0.28	0.668
[ethnic_names_with_prior_performance_boy_South Asian]																-0.21	-0.57–0.15	0.245
Random Effects																		
σ^2	0.95			1.10			1.46			0.77			0.82			1.48		
τ_{00}	1.46	problem_id		0.97	problem_id		0.98	problem_id		1.67	problem_id		1.49	problem_id		1.38	problem_id	
ICC	0.22	teacher_xid		0.54	teacher_xid		0.17	teacher_xid		0.21	teacher_xid		0.33	teacher_xid		0.06	teacher_xid	
N	0.64			0.56			0.44			0.71			0.69			0.49		
	18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid	
	36	problem_id		32	problem_id		33	problem_id		31	problem_id		36	problem_id		38	problem_id	
Observations	360			360			360			360			360			360		
Marginal R ² / Conditional R ²	0.000 / 0.638			0.001 / 0.562			0.003 / 0.441			0.001 / 0.711			0.002 / 0.690			0.002 / 0.493		

Table A.12: Comparing sub-effects across conditions for different ethnicities separately for girls.

Predictors	grade			grade			grade			grade								
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p						
(Intercept)	2.51	2.12 – 2.91	<0.001	2.16	1.68 – 2.65	<0.001	1.72	1.22 – 2.22	<0.001	2.31	1.89 – 2.73	<0.001	1.88	1.38 – 2.38	<0.001	2.14	1.60 – 2.68	<0.001
[anonymized_with_prior_performance_girl_African American]	-0.08	-0.32 – 0.15	0.487															
[ethnic_names_no_prior_performance_girl_African American]	0.03	-0.21 – 0.26	0.817															
[ethnic_names_with_prior_performance_girl_African American]	0.01	-0.23 – 0.24	0.963															
[anonymized_with_prior_performance_girl_Asian]				-0.01	-0.34 – 0.32	0.947												
[ethnic_names_no_prior_performance_girl_Asian]				0.07	-0.26 – 0.40	0.690												
[ethnic_names_with_prior_performance_girl_Asian]				0.09	-0.24 – 0.42	0.595												
[anonymized_with_prior_performance_girl_Caucasian]							-0.03	-0.27 – 0.21	0.822									
[ethnic_names_no_prior_performance_girl_Caucasian]							0.01	-0.24 – 0.25	0.964									
[ethnic_names_with_prior_performance_girl_Caucasian]							0.06	-0.19 – 0.30	0.652									
[anonymized_with_prior_performance_girl_Hispanic]							-0.04	-0.30 – 0.22	0.767									
[ethnic_names_no_prior_performance_girl_Hispanic]							-0.06	-0.32 – 0.20	0.641									
[ethnic_names_with_prior_performance_girl_Hispanic]							-0.03	-0.29 – 0.22	0.799									
[anonymized_with_prior_performance_girl_Middle Eastern]										0.12	-0.18 – 0.42	0.427						
[ethnic_names_no_prior_performance_girl_Middle Eastern]										0.22	-0.08 – 0.52	0.149						
[ethnic_names_with_prior_performance_girl_Middle Eastern]										0.10	-0.20 – 0.40	0.516						
[anonymized_with_prior_performance_girl_South Asian]																		
[ethnic_names_no_prior_performance_girl_South Asian]																		
[ethnic_names_with_prior_performance_girl_South Asian]																		
Random Effects																		
σ^2																		
τ_{00}	1.29			1.26			1.37			1.55			1.06			0.69		
ICC	0.70	problem_id		1.16	problem_id		1.14	problem_id		0.81	problem_id		1.23	problem_id		1.31	problem_id	
	0.24	teacher_xid		0.19	teacher_xid		0.41	teacher_xid		0.21	teacher_xid		0.13	teacher_xid		0.44	teacher_xid	
	0.42			0.52			0.53			0.40			0.56			0.72		
	18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid		18	teacher_xid	
	47	problem_id		38	problem_id		46	problem_id		42	problem_id		30	problem_id		38	problem_id	
Observations	720			360			720			720			360			360		
Marginal R ² / Conditional R ²	0.001 / 0.421			0.001 / 0.518			0.000 / 0.532			0.000 / 0.396			0.003 / 0.562			0.001 / 0.716		

Chapter B

SUPPLEMENTARY MATERIALS FOR CHAPTER 11 “LIVE-CHART: LIVE INTERACTIVE VISUAL ENVIRONMENT FOR CREATING HEIGHTENED AWARENESS AND RESPONSIVENESS FOR TEACHERS ”

B.1 Task Abstraction

The development of goals and task abstraction for TA was an iterative process. In order to understand the requirements of a TA tool, we interacted with various domain experts, experienced teacher trainers, teachers, and researchers to understand the fundamental goals a TA tool needs to accomplish. Additionally we also analyzed the four TA tools presented in figure 11.2 as they provide valuable insight into the teacher needs and how to best address them. We divided the overall analysis into two parts. We began with goals analysis, where we developed a hierarchy of goals that a TA tool needs to facilitate. Next, we deconstructed the goals into subgoals. Subgoals give us high-level objectives that directly correlate with teacher needs. We leverage the concrete set of subgoals to enumerate visualization tasks in the TA tool that directly facilitate the teacher’s needs. We went through multiple iterations of goals and tasks analysis to further refine our findings.

Through LIVE-CHART, we have expanded upon the affordances of Teaching Augmentation systems to facilitate an effective student-teacher interface. LIVE-CHART provides teachers with insights into their students’ performance on classwork, which influences the quality of student-teacher interactions as teachers address students’ needs.

B.1.1 Goal Analysis

Table B.1 lists the goals and subgoals resulting from our analysis. The overarching goal of TA is to augment teachers’ ability to interface with their students at an individual and class level. The goals help teachers identify the instances where they can address students one-on-one vs. as a group vs. the entire class. The first three goals, G1, G2, and G3, directly address teacher needs from various perspectives. The teachers need to analyze the entire class at a glance. Identify attendees, absentees, students who completed their classwork, and students who left their classwork incomplete. The teachers, at a glance, should be able to identify assignment progress for the entire class and individual students. When teachers are interested in comprehending the underlying causes behind student performance, TA should help teachers gain contextual insight. Infer the effort and attention a student put towards solving a problem and how close they came to solving it. Teachers have often expressed a wish to clone themselves so that they can help their students more effectively. TA helps teachers augment their abilities by helping them identify the students who require attention or students doing well. Goal 3 facilitates teachers to provide positive affirmations to students doing well and help students who require attention. Providing quantitative information on student performance is not enough as the teachers know their students and make nuanced inferences that a TA cannot make. For instance, student A taking their time working on the assignment might mean they need help, whereas student B taking the same amount of time might mean they are just thorough. These are nuances that come easily to a teacher but are challenging to quantify in TA.

B.1.2 Task Analysis

We identified the goals in Table B.1 related to abstract visualization tasks from Brehmer and Munzner’s topology: the high-level family of Present tasks and the lower level family of Browse, Explore, and Identify tasks [49].

B.2 Custom Seating Arrangement

B.3 Alternative Seating Arrangements

Figure B.2 is the visualization of students in an alphabetical order and figure B.3 is the anonymized version of the alphabetical arrangement. Similarly, Figure B.4 is the visualization of students in an per-problem view and figure B.5 is the anonymized version of the per-problem view.

Table B.1: Fundamental goals of a TA tool.

Generic Goals	
G1	Analyze assignment progress
a	Analyze assignment(classwork) progress for the entire class
b	Analyze assignment(classwork) progress for individual students
c	Identify problems where students struggled the most
G2	Identify underlying causes of differences in student performance
a	Use the student actions while working on the problems to gain contextual insight into student performance and effort
b	Infer the amount of effort and attention a student put in towards solving the problem
c	Identify problems where students struggled the most
G3	Enhance teacher performance
a	Identify students who are doing well or require attention
b	Help teachers facilitate equity in the classroom
G4	Discover nuances (Qualitative Inferences the teacher can make off of the quantitative data)

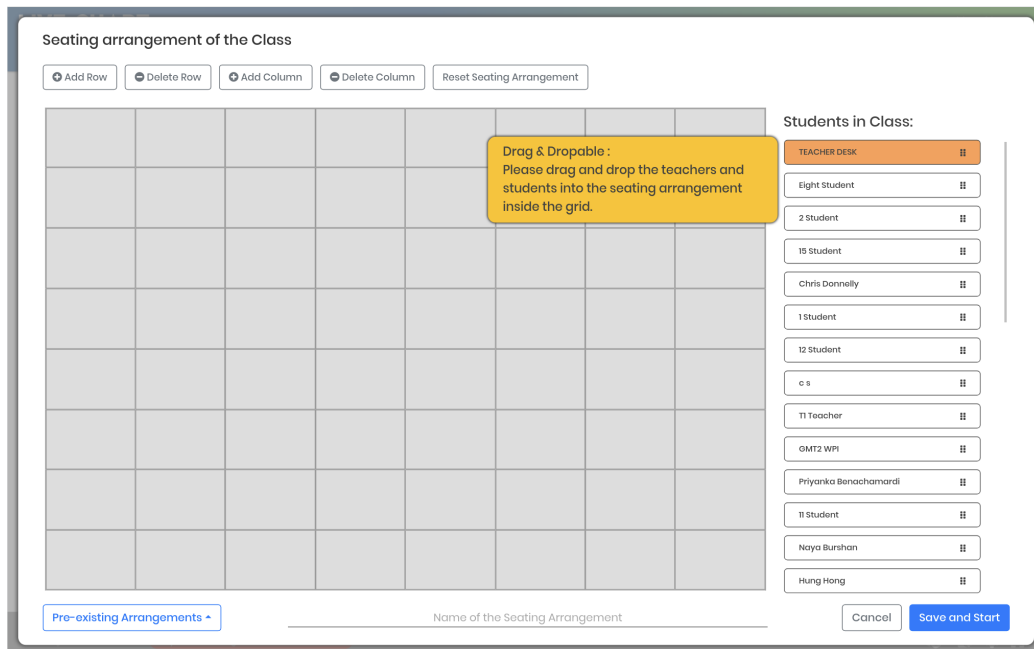


Figure B.1: Seating Arrangement: teachers can arrange the students in the class to reflect the seating arrangement of the class.

Table B.2: Fundamental goals of a TA tool.

Tasks	
G1. a. Analyze assignment(classwork) progress for the entire class	
T1	Identify students who are absent and students who are present.
T2	Differentiate students who are working on the assignment, who have completed it, and who left without completing the assignment
T3	Analyze the difference between students who are quick vs. the students who take their time working on the assignment.
G1. b. Analyze assignment(classwork) progress for individual students	
T4	Analyze the problem correctness for the recent problems the student worked on
T5	Identify the percentage of the assignment the student completed
T6	Identify if the student is absent, working on the assignment, completed the assignment, or left without completing the assignment
G2. a. Use the student actions while working on the questions to gain contextual insight into student performance	
T7	Identify the number of attempts a student made at answering
T8	Identify the usage of Hints and explanations
T9	Identify the time on task (amount of time students spent on the problem)
T10	Identify if the students asked for the answer
T11	Understand the approach the student took to solve the problem
G2. b. Infer the amount of effort and attention a student put in towards solving the problem	
T12	Identify the amount of time a student is trying to solve the problem
T13	Identify the amount of time a student spent understanding the help provided by the system. E.g. Hints, explanations
G2. c. Identify how close students were to answer the problem	
T14	Infer the approach the student took to solving the problem
T15	Identify the number of attempts the students made to solve the problem
T16	Infer how close the students were to the answer E.g. common wrong answer, a common misconception, silly mistakes
G3. a. Identify students who are doing well or require attention	
T17	Help direct the teacher attention to students who need help
T18	Create opportunities for the teacher to provide positive feedback to students who are doing well
G3. b. Help teachers facilitate equity in the classroom	
T19	Help teachers identify growth or decline in students' performance in a manner that is independent of the teachers' perception of student abilities
T20	Help improve the quality of student-teacher interaction by helping the teacher infer the approach and effort the student invested in answering problems

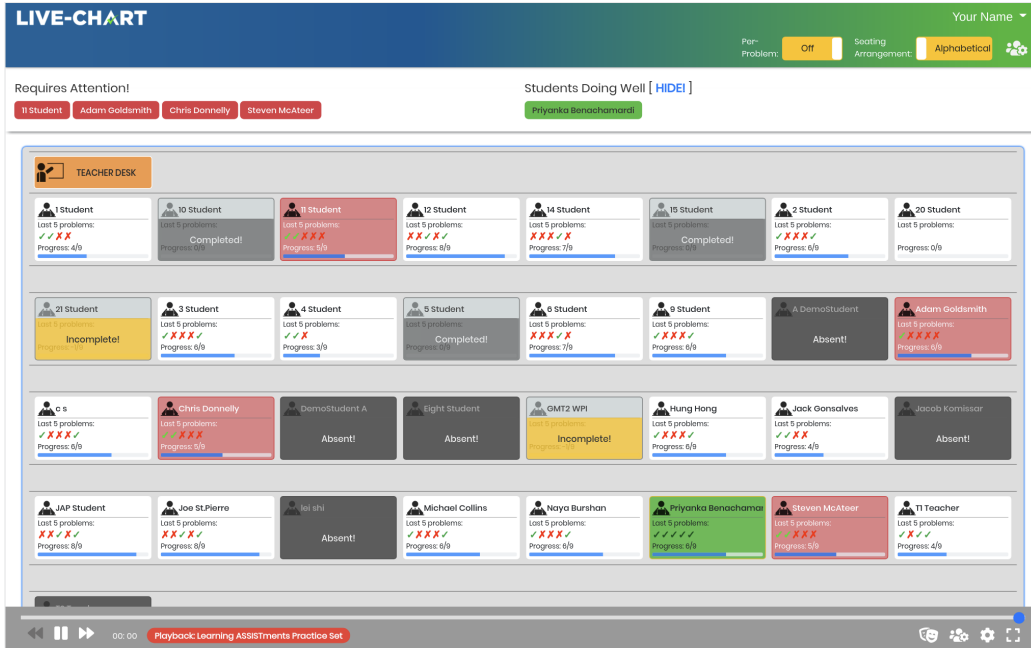


Figure B.2: Visualization of students in Alphabetical order.

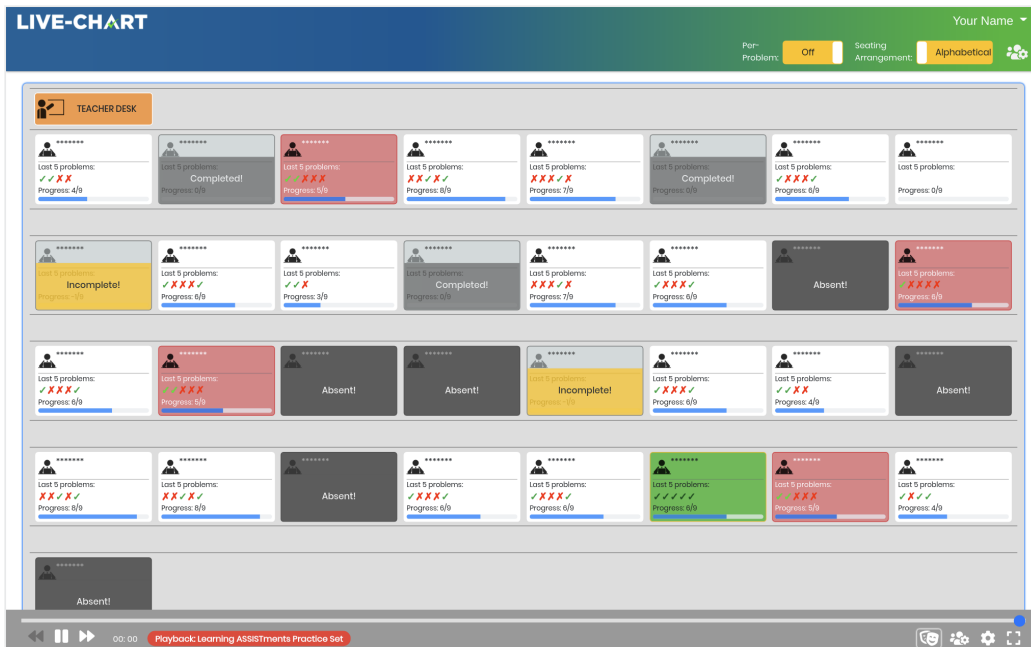


Figure B.3: Visualization of students in anonymized Alphabetical order.

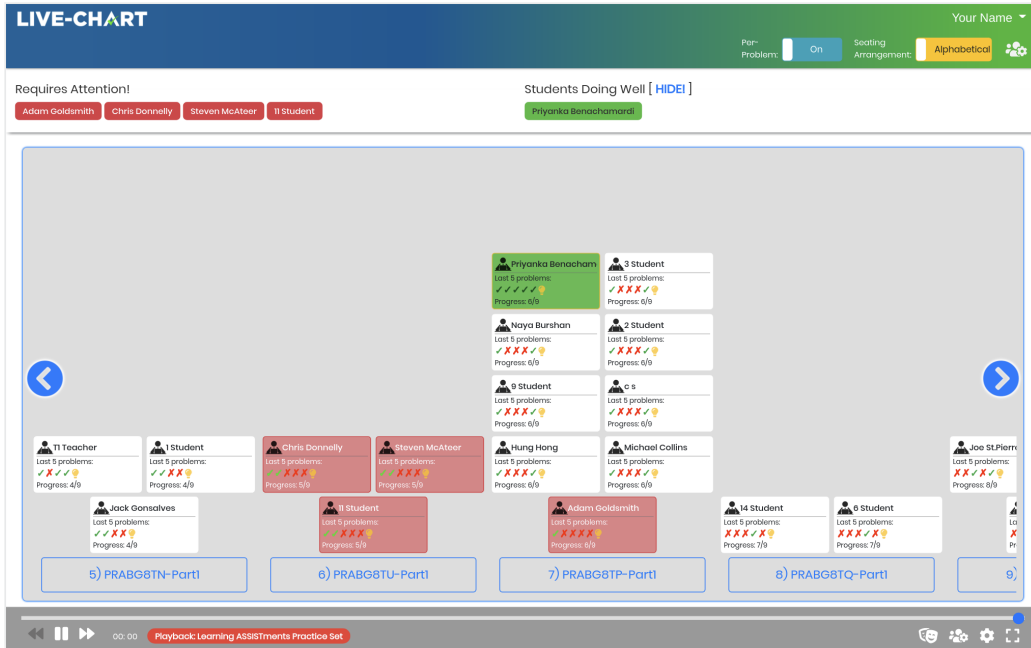


Figure B.4: Visualization of students in per-problem view.

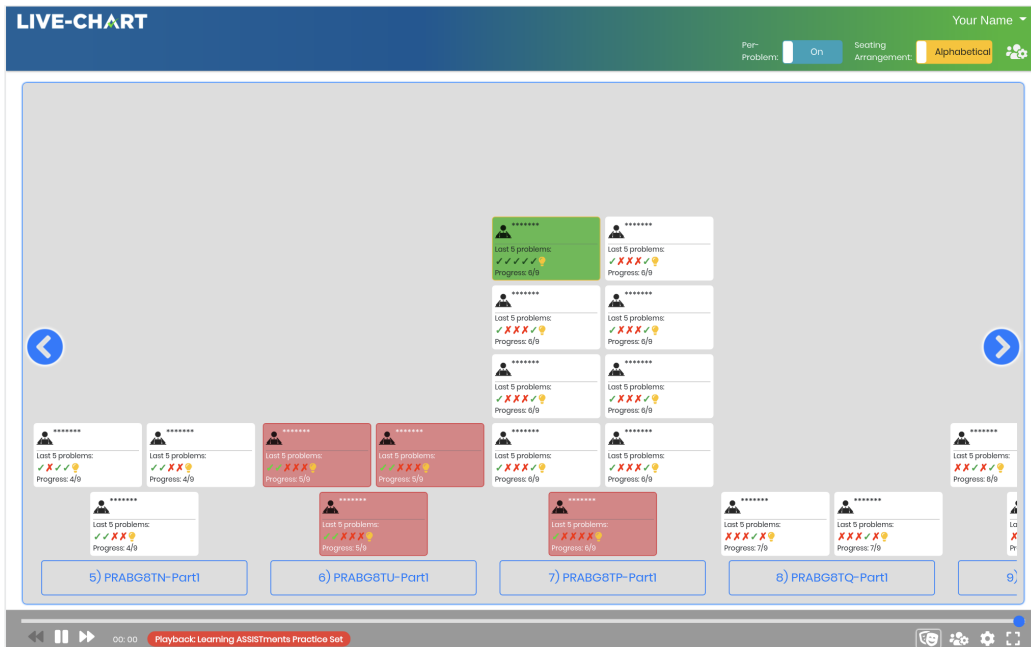


Figure B.5: Visualization of students in anonymized per-problem view.