Toward Improving Effectiveness of Crowdsourced, On-Demand Assistance From Authors in Online Learning Platforms

by

Aaron Haim

A Thesis Proposal

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master in Science

in

Computer Science

April 2022

Approved:

Professor Neil T. Heffernan, Thesis Advisor

Professor Jacob R. Whitehill, Thesis Reader

Professor Craig A. Shue, Head of Department

Abstract

In this experiment, we have a set of **authors** made up of teachers and undergraduate college students who we paid to write *student-supports*, which are typically hints and explanations, to be given to students on-demand while solving problems assigned by their teachers in the ASSISTments platform[1]. We want to see if we can tell which authors are, on average, producing *student-supports* that cause better student learning. We conducted a month-long intervention where students were exposed to support from different authors. In this experiment and its replication, we randomized the authors of the *student-support*s and analyzed a set of pairwise comparisons between authors. We failed to find evidence that we can reliably tell the difference between authors. It could be that our authors produce equally effective *student-supports*, or it could be that this work was underpowered, and we failed to recruit enough students to discover existing differences. All data and analysis being conducted can be found on the Open Science Foundation website[2].

[1] https://www.assistments.org
[2] https://osf.io/zcbjx/

Acknowledgements

We would like to thank the NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), and Schmidt Futures. None of the opinions expressed here are that of the funders. We are funded under an NHI grant (R44GM146483) with Teachly as a SBIR.

Table of Contents

Abstract ii

Acknowledgements iii

Table of Contents iv

Illustrations v

Nomenclature vi

Introduction 7

Background 7

Methodology 9

Analysis 11

Results 13

Conclusion 15

References 17

- Appendix 17
  - Appendix A: RQ1 Results for Study 1 and 2 17
  - Appendix B: RQ2 Results for Study 1 and 2 25
  - Appendix C: Author Code to Author Identifier 33
  - Appendix D: Consort Data Flow Plan for Dataset A 34
- Bibliography 35

Illustrations

Fig. 1: A set of hints (Left) and an explanation (Right) for the sample problem in the ASSISTments platform. 8

Table 1: Timeline of the study breakdown of the data collection and method used to select the student-support to deliver to the student on request. 9

Fig. 2: The percentage of *student-supports* each author has generated within the ASSISTments platform. 11

Fig. 3: Model for the main effect, Dataset A Study 1 (the feature "author" is a categorical variable representing the first author in the *student-supports* author pairs i.e., author A). 13

Fig. 4: A matrix of all possible pairwise comparisons between any two authors for Study 1 (left) and Study 2 (right). Author identifiers can be translated using Appendix C. 14

Table 2: Study 1 overview of main effects. 15

Table 3: Study 2 overview of main effects. 15

Nomenclature

- **Author:** A creator of *student-supports* within the ASSISTments platform.
- ***Student-Support*:** A piece of feedback created by an author, typically a hint or explanation.
- ***Star-Author:*** An author whose *student-supports* can be seen by any student in the ASSISTments platform.
- **Single-Support Randomization:** A randomization method that occurs when only one *student-support* can be selected from for a problem. 90% of the time, the *student-support* can be requested by the student. The other 10%, only the answer can be requested by the student.
- **Problem-Based Randomization:** A randomization method that occurs when only multiple *student-supports* can be selected for a problem. A *student-support* is randomly selected from a list of *student-supports* available.
- **Author-Based Randomization:** A randomization method that occurs when multiple *student-supports* can be selected for a problem. A *student-support* is selected according to a priority list of *star-authors* assigned to the student.
- **Next Problem Correctness:** A boolean dependent measure used in previous works that is true if the student answered the next problem after receiving a *student-support* correct on the first try without viewing another *student-support*.
- **Dataset A:** The initial dataset containing the data for the two authors used to select the features for the OLS model.
- **Dataset B:** The main dataset containing the data for all remaining pairwise comparisons of authors.

Introduction

Studies have proven that providing on-demand assistance and additional instruction on a problem when a student requests it improves student learning in online learning environments. Additionally, crowdsourced, on-demand assistance generated from authors in the field is also effective. However, these studies conduct problem-based randomization where each condition represents different student-support for every problem encountered. As such, claims about a given author's effectiveness are provided on a per-student-support basis and not easily generalizable across all students and problems.

The ASSISTments project is trying to be the premier digital platform that supports high-quality studies in authentic, digital classroom environments. We can do this because thousands of teachers use ASSISTments to assign their classwork and homework, and we design numerous randomized controlled trials to learn what helps students learn. The science on the principles of learning always has a give and take between collecting observation data, engaging in theory building, and mixing in some amount of experimentation. For instance, we can take the principles to design and execute experiments to study their effect on learning.

Experiments using theory have a role to play in science. They generally manipulate a single variable simultaneously and help build new theories. But there is also a role of observing what works and then hypnotizing why something might be working. In this experiment, we are trying to see if we can detect a reliable difference in student learning between author. After we do that, we should be left with a set of content from authors that work well and a set of supports that don't work as well, and we can hypothesize what features of authors' *student-supports* are most effective and then use the E-TRIALS infrastructure to build two sets of student supports that differ only by that feature (Krichevsky, 2020).

As such, this experiment aims to answer the following questions:

RQ1 When comparing two authors, which is the most effective at generating *student-supports (i.e., who causes the biggest gain on the post-test)*?

RQ2 When comparing authors, are there reliable differences based upon demographics? (i.e., do lower knowledge students perform better with *student-supports* generated by author X compared to Y? Does one author write feedback that is better for females? Does one author write feedback that is better for students in rural schools?)

Background

As online learning platforms expand their content base, the need to generate on-demand assistance grows alongside it (Patikorn & Heffernan, 2020). Crowdsourcing provides an effective method to generate new assistance for students (Heffernan & Heffernan, 2014). As on-demand assistance generally improves student learning, authors and their assistance must be evaluated to maintain or improve the current level of quality of effectiveness (McLaren et al., 2016; Razzaq & Heffernan, 2009; Wood et al., 1976).

In 2003, Neil T. Heffernan and Cristina Heffernan developed ASSISTments: a free, online learning platform providing feedback and insights on students to better inform teachers for classroom instruction (Heffernan & Heffernan, 2014). ASSISTments provides problems and assignments from open source curricula, the majority of which is K-12 mathematics, which teachers can select and assign to their students. Students complete assigned work within ASSISTments. For most problem types, students receive immediate feedback when a response is submitted for a problem, which tells the student whether the answer is correct and, if not, allows the student to try again (Feng & Heffernan, 2006).

In 2017, ASSISTments deployed the Special Content System, formerly known as TeacherASSIST. Dr. Heffernan had met teachers who were writing hint messages for their own students, but Heffernan had not built support for this function into the platform, preventing other teachers from assigning these author-created messages. This new system we created allows authors whom we trust to have their content go "viral" across the system. The new system called the "Special Content System" allows authors to create on-demand assistance or ***student-supports*** within the platform. This allowed us to identify which authors are making good content.

When ONLY ONE *student-support* was available for a given problem, the Special Content System performed a **single-support randomization**, where a given student would have a 90% chance of receiving the *student-support* with a 10% chance of receiving no *student-support*. Single-support randomization was evaluated based on the student's ability to answer the next problem correctly on the first try, known as **next problem correctness**. Using single-support randomization, we found that delivering *student-supports* to students caused more student learning compared to immediately giving students the answer (Patikorn & Heffernan, 2020; Prihar et al., 2021).

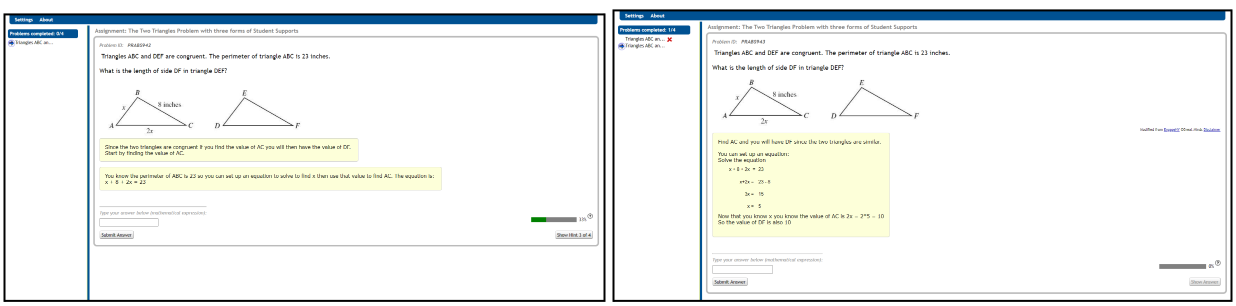


Fig. 1: *A set of hints (Left) and an explanation (Right) for the sample problem in the ASSISTments platform.*

When TWO OR MORE *student-supports* were available for a given problem, the Special Content System performed a **problem-based randomization**, where a given student would be randomly assigned one of the available *student-supports*. Using problem-based randomization, we were able to assess which authors were more effective at improving student learning compared to other authors (Prihar et al., 2021). As such, claims about a given author's effectiveness are provided on a per-*student-support* basis– but we still don't know which author was generally better at improving student learning. In addition, students learn information cumulatively across problems (Lee, 2012), making it difficult to generalize this claim across all, or at least certain subsets, of students and problems within the platform.

The data ASSISTments collects from these various random control trials are highly valuable. We examined overall trends across various experiments, presenting the results of 50+ experiments involving over 50,000 students that tested many different ideas, including 1) giving student choices, 2) motivational messages, and 3) fill-in-the-blank versus multiple choice (Prihar, Syed, et al., 2022). We failed to find a main effect of giving students choices and surprisingly found that giving motivational messages backfired and was associated with poorer performance. Finally, we found that fill-in-the-blank answer types caused reliably better student learning.

As the ASSISTments platform determines which *student-supports* for a given problem are the most effective at improving student learning in general, there has been additional research to personalize which *student-supports* are better for a given student (Prihar et al., 2022)--shifting from problem-focused support to student-focused support. If ASSISTments chooses to develop a personalized learning approach for delivering *student-supports* to students, then it would be more difficult for the platform to evaluate new *student-supports* or authors without negatively impacting a student's learning. For example, let's say that for 40 students, we know which *student-support* for a specific problem will improve their performance the greatest. If another author added a new *student-support* for the given problem, we would have a high potential to detriment the students' learning without any prior data about whether the given *student-support* or any of its contributing factors are effective. By evaluating the general effectiveness of an author, new *student-supports* from effective authors could be introduced into the personalization model without majorly disrupting a student's learning. In addition, new students may receive *student-supports* more often from a given author in addition to the most effective *student-support* written for a problem to more efficiently determine which *student-support* would be more effective for a particular student.

Methodology

This experiment modified the Special Content System to use either problem-based randomization or author-based randomization over the course of three-and-a-half months. During this period, the initial study, known as **Study 1**, and a replication study, known as **Study 2**, delivered *student-supports* to students via author-based randomization across *star-authors* for the course of a month. To measure the performance of a given student, there was a two-week interval before Study 1, known as the **Pre-Test**, a two-week period in-between Study 1 and Study 2, known as the **Mid-Test**[3], and a two-week period after Study 2, known as the **Post-Test**. During the test phases, we still gave students *student-supports*; it was just random. The tests will be treated as the initial state and the dependent measure to determine a student's growth in learning during the period of the author-based randomization.

[3] The Mid-Test will act as a posttest to Study 1 and a pretest to Study 2.

Phase	Length of Time	Selection Mechanism
Pre-Test	2 Weeks Feb 16, 2022, to Feb 28, 2022	Problem-Based Randomization
Study 1	1 Month March 1, 2022, to March 31, 2022	Author-Based Randomization
Mid-Test	2 Weeks April 1, 2022, to April 15, 2022	Problem-Based Randomization
Study 2	1 Month April 16, 2022, to May 15, 2022	Author-Based Randomization
Post-Test	2 Weeks May 16, 2022, to June 1, 2022	Problem-Based Randomization

Table 1*: Timeline of the experiment breakdown of the data collection and method used to select the student-support to deliver to the student on request.*

Study 1: Author-Based Randomization

Study 1 will use author-based randomization over a period of a month. Ideally, every student could be assigned to a particular star-author. However, authors have the choice to write one student-support per problem for any problems they wish. As such, star-authors can generate student-supports across any number of problems with as much or as little overlap with other star-authors. As shown in Fig. 2, in ASSISTments, twenty star-authors have collectively generated 53,817 student-supports; however, four star-authors have generated over 50% of the available student-supports, with only two generating above 10% of the total pool.

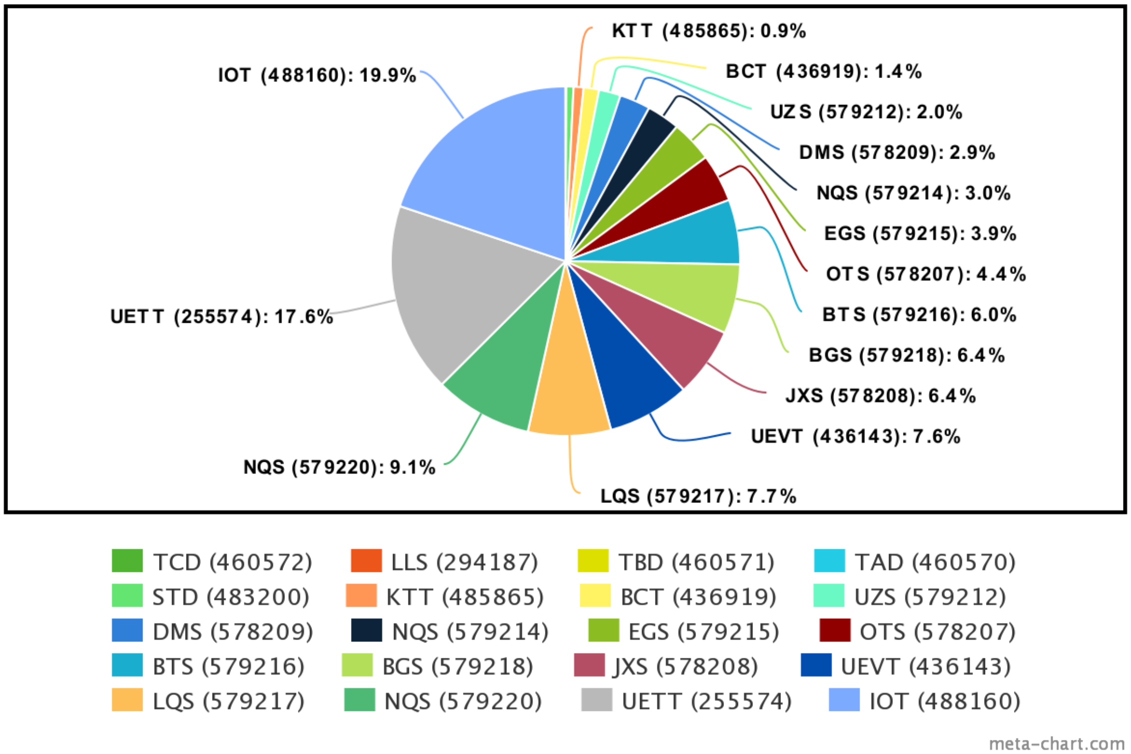


Fig. 2*: The percentage of student-supports each author has generated within the ASSISTments platform.*

If an author has not written a student-support for the problem the student is solving and another author has, the Special Content System should provide one of the available student-supports. As such, assigning a single author to a given student would prevent students from receiving student-supports from authors who wrote a small number of them.

To mitigate the issue, a random ordering of all available star-authors was assigned to each student across the experiment period. This allowed a student to remain in condition with a given author for as long as possible. When a student requested a student-support for a given problem, the student will receive a student-support from the **topmost** author in their list ordering, who has written a student-support for a problem. For example, if there are three authors in the ordering B, A, and C, we first examine whether author B has written a student-support for that problem. If not, we examine author A and so on until an author has written a student-support for the problem or there are no student-supports.

Study 2: Author-Based Randomization with Reversed Ordering

Study 2 will also use an author-based randomization following the two-week interval known as the Mid-Test. Compared to Study 1, students will be provided a student-support from the **bottommost** author in their list ordering, which has written a student-support for a problem. Using the previous example with the author ordering B, A, C, we first examine whether author C has written a student-support for that problem. The Mid-Test will be treated as the pretest for Study 2 to account for the changes in performance gained across Study 1.

Power Analysis

We conducted a power analysis in R using the pwr package (Team, R.C., 2013; Champely, 2020), assuming that our intervention would double the normative expectation of change. Lipsey et al. (2012) suggest using a standardized instrument for the normal amount of change for 7-8th graders, which corresponds to an effect size of d = 0.32. To achieve 80% power, with alpha = 0.05, we will need a total of n = 310 students.

Analysis

We preregistered the conditions in our experiment, but since we had not written any analysis code at the time, we stated in our first pre-registration that we would pull down a small sample of about 10% of the collected data, then use that to create an analysis plan to analyze the remaining 90% of the data (we call the first dataset, **Dataset A** and we call the primary dataset, **Dataset B**). After using Dataset A, we will never again look at Dataset A.

To be precise, our criterion for analysis was two-fold: 1) that at least 1,000 students were exposed to a randomized controlled comparison between a given pairwise group of authors and 2) since we did not want to confound the type of student support (whether they wrote hints or an explanation) we only wanted to compare authors who wrote the same type of supports. Based on prior months, only 35% of students requested a student-support. As such, since we tried to observe as little data as possible, we calculated that each pairwise comparison should have at least 886 students. We then rounded-up the value to 1,000 students to account for potentially lost data and overlap between conditions.

Inclusion Criteria

To generate the initial model, we used Dataset A, allowing us to solidify the method for handling Dataset B. Dataset A included students who viewed a problem during the Study 1 time period, and they requested a *student-support* and could have received either author in the pair BCT (436919) and EGS (579215)[4]. In the case of three or more author conditions, the student had to be randomly assigned to one of the authors in the comparison (e.g., BCT or EGS). We then looked at the two-week period prior to the study, referred to as the Pretest. Students who did not complete at least one problem during the Pretest period were excluded from the study. Similarly, students who were not assigned any problems during the study posttest period were excluded from the study.

For the BCT vs EGS comparison, 1,073 students met the initial eligibility requirements. 345 students did not complete any pretest problems and were excluded leaving 728 students randomized between BCT and EGS: 373 to BCT and 355 to EGS. In the BCT condition, 316 students were excluded: 305 did not ever request a *student-support* during the study period, and 11 more were not assigned a problem during the post-test period. In the EGS condition, 297 were excluded, 282, due to not ever requesting a *student-support,* and 15 were not assigned a problem during the post-test. (To be clear, if a student never asks for *student support*, they have no idea what condition they would have gotten, so it's very reasonable to drop all students who never requested a *student-support*.) This left 57 students in the BCT condition and 58 students

[4] A breakdown of which Author Code belongs to which Author Identifier can be found in Appendix C.

in the EGS condition to generate the model. We used this pair to create the model, but since the total number of students was under 310, it would not pass the criteria for our power analysis. But that was the point: pull a small amount of data to make Dataset A that we can use to write a precise data filtering and analysis plan to preregister. The flow diagram showing this enrolment cascade is in Appendix D.

The Preregistration of Analysis Plan using Dataset A

For each student, we collected statistics prior to the experiment period, the author condition they were assigned to during the course of the study, and the average partial credit score across all problems on the pretest and posttest. We then used the statistics, author condition, and average partial credit score on the pretest as the initial feature set to fit an Ordinary Least Squares (OLS) mode and observe the exact coefficient on the author condition. The average partial credit score on the posttest acted as the dependent measure.

Using the initial feature set and the analysis model, we first screened features for collinearity. If the correlation between a pair of features was greater than 0.95 in absolute value, one feature of the pair (chosen arbitrarily) was dropped. Afterward, the remaining features in the model were removed one at a time using a backward stepwise regression. The regression would remove the feature that was the most insignificant. The author condition and average partial credit score across the pretest were static features and were not removed from the model (using our step-wise process). The remaining features were then fixed in the model to generate the interaction effects and removed high correlations and insignificant ones.The model is shown in Fig. 3.

```
                              OLS Regression Results
==================================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.389
Model:                                  OLS   Adj. R-squared:                  0.367
Method:                       Least Squares   F-statistic:                     24.43
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           1.75e-14
Time:                              11:10:47   Log-Likelihood:              -0.67835
No. Observations:                       115   AIC:                             11.36
Df Residuals:                           110   BIC:                             25.08
Df Model:                                 4
Covariance Type:                        HC1
==================================================================================================
                                         coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                                  0.6843      0.092      7.452      0.000       0.504       0.864
author                                -0.0544      0.047     -1.163      0.245      -0.146       0.037
pretest_avg_problem_accuracy           0.2831      0.094      3.020      0.003       0.099       0.467
student_std_attempted                  0.5912      0.262      2.258      0.024       0.078       1.104
student_std_attempted_before_support  -1.0845      0.221     -4.914      0.000      -1.517      -0.652
==============================================================================
Omnibus:                        0.859   Durbin-Watson:                   2.303
Prob(Omnibus):                  0.651   Jarque-Bera (JB):                0.965
Skew:                          -0.187   Prob(JB):                        0.617
Kurtosis:                       2.752   Cond. No.                         16.4
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

***Fig. 3**: Model for the main effect, Dataset A Study 1 (the feature "author" is a categorical variable representing the first author in the Student-Supports author pairs i.e., author A).*

We "burned" (i.e., we used some data to generate an analysis plan that we then never used again) this one pairwise comparison (BCT vs. EGS) to write code to analyze the other pairwise comparisons in Study 1 and Study 2. To avoid p-hacking, we ran the analysis a single time, only touching the data once, to generate the necessary results.

The Main Dataset: Dataset B

The selected features were then used to fit an OLS model for the remaining pairwise comparisons of author-pairs. If the author-pair condition was significant and the confidence interval did not include zero, then we could claim that one author outperformed the other. The author's condition was significant if the p-value, corrected by Benjamini-Hochberg, was less than 0.05.

Demographic Results

In addition to the author condition model, a separate OLS model was fitted with the selected features and three demographic features along with their interactions with the author condition. The demographic features collected were the inferred gender of the student, whether the school the student attended was in an urban, suburban, or rural setting, and whether the student was in the top third, middle third, or lower third of students based on the average partial credit score across the posttest.

Results

After running the analysis, Study 1 only had nine pairwise comparisons that met the initial inclusion criteria, while Study 2 only had five. This can be seen in Fig. 4, where the nine red boxes show the valid pairwise comparisons. The summarized results of Study 1 and Study 2 which met the inclusion criteria can be seen in Table 2 and Table 3, respectively. The full results can be found in Appendix A.

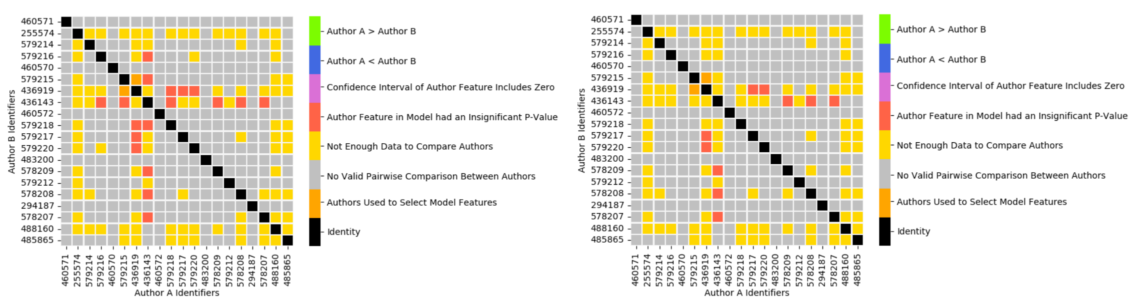


Fig. 4: *A matrix of all possible pairwise comparisons between any two authors for Study 1 (left) and Study 2 (right). Author identifiers can be translated using Appendix C.*

The power analysis suggested that we needed 310 students, so we only looked at experiments where the number of students was over 310. However, since 1) many of the students did not complete a problem during the two-week pretest period and 2) only 35% of students asked for a *student-support* during the month of the study, many of the pairwise comparisons did not have enough students to be considered significant. Out of the nine comparisons on Study 1, only three have over 310 students, and none of those experiments suggested a difference between authors. In Study 2, we found only one student with over 310 students, and that study also failed to find a main effect between authors.

Pairwise Group ID (Author A vs. B)	P-Value (Corrected)	Effect Size Estimate	Number of Observations	Meet Power Analysis (n>310)	Demographic
UEVT OTS			286	No	
UEVT JXS	.274 (0.739)	.0281	420	Yes	No reliable interactions.
UEVT DMS			245	No	
UEVT EGS			165	No	
UEVT BTS			125	No	
UEVT BGS			129	No	
BCT _LQS	.106 (.739)	.0367	533	Yes	No reliable interactions.
BCT BGS			247	No	
BCT _NQS	.562 (.778)	-.0159	361	Yes	No reliable interactions.

***Table 2**: Study 1 overview of main effects.*

Pairwise Group ID (Author A vs. B)	P-Value (Corrected)	Effect Size Estimate	Number of Observations	Meet Power Analysis (n>310)	Demographic
UEVT OTS			101	No	
UEVT JXS		.	114	No	
UEVT DMS			123	No	
BCT LQS	.340 (.965)	.0268	364	Yes	A reliable interaction.
BCT NQS			242	No	

***Table 3**: Study 2 overview of main effects.*

For RQ2, we looked to see if the four comparisons had reliable interactions with demographic features and conditions. We found that in Study 1, across the three comparisons, there were none (summarized in the right column of Table 2), while in Study 2, only one comparison found a reliable effect of locale. This was interpreted to mean that for students in a school located in an urban district, they performed reliably better with one of the authors. However, given that in Study 1, we did not find that effect for the same pair of authors (436919 - 579217), so we are not making much of that finding. The full results can be found in Appendix B.

Conclusion

In this experiment, using the dynamically selected model, we failed to find evidence that we can find reliable differences in student learning. That does not mean there is no difference and authors are equally good; we can only conclude that this plan failed to find differences. A couple of significant differences were found within the demographic model, but they are likely to be attributed to the variance of the feature set. We could try to run a planned comparison to see if those interactions could be replicated.

Limitations

We had 140,365 students use ASSISTments since July 1, 2021. We had 32,057 middle school students using ASSISTments during the period of the study, but we are reporting on experiments with just hundreds of students. We were surprised that the n-sizes in our experiments were so small. But we wrote ahead of time a detailed pre-registration specifying who qualified to participate. Since we only allowed students that attempted one problem during the pretest period, we lost subjects. They also had to ask for a *student-support* during the study. Therefore we lost many users as they never asked for help (so they never saw the conditions).

We also suffered from having 20 different authors write content, so there were too many author-pair conditions to have a lot of subjects per condition. One thing we want to change in this next round is to get more statistical power to detect differences. In this past study, the students were divided into many different conditions making the total for each condition lower than we would have liked.

References

Appendix

Appendix A: RQ1 Results for Study 1 and 2

This section shows the nine different regression results for Study 1 for models with main effects. These are the models that relate to RQ1. Table 2 summarized a few key results from the below nine regressions.

i. UEVT (436143) vs OTS (578207)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.241
Model:                                  OLS   Adj. R-squared:                  0.230
Method:                       Least Squares   F-statistic:                     20.06
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           1.51e-14
Time:                              11:10:49   Log-Likelihood:                -1.8209
No. Observations:                       286   AIC:                             13.64
Df Residuals:                           281   BIC:                             31.92
Df Model:                                 4
Covariance Type:                        HC1
==========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                    0.7449      0.080      9.311      0.000       0.588       0.902
author                                  -0.0150      0.029     -0.517      0.605      -0.072       0.042
pretest_avg_problem_accuracy             0.1845      0.082      2.261      0.024       0.025       0.345
student_std_attempted                   -0.5110      0.247     -2.070      0.038      -0.995      -0.027
student_std_attempted_before_support    -0.6978      0.194     -3.598      0.000      -1.078      -0.318
==============================================================================
Omnibus:                       22.872   Durbin-Watson:                   1.691
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.590
Skew:                          -0.659   Prob(JB):                     1.68e-06
Kurtosis:                       3.703   Cond. No.                         20.2
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

ii. UEVT (436143) vs JXS (578208)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.207
Model:                                  OLS   Adj. R-squared:                  0.200
Method:                       Least Squares   F-statistic:                     28.45
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           6.58e-21
Time:                              11:10:51   Log-Likelihood:                -35.559
No. Observations:                       420   AIC:                             81.12
Df Residuals:                           415   BIC:                             101.3
Df Model:                                 4
Covariance Type:                        HC1
==========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                    0.5664      0.060      9.488      0.000       0.449       0.683
author                                   0.0281      0.026      1.093      0.274      -0.022       0.079
pretest_avg_problem_accuracy             0.2989      0.056      5.351      0.000       0.189       0.408
student_std_attempted                   -0.0981      0.183     -0.536      0.592      -0.457       0.261
student_std_attempted_before_support    -0.6172      0.148     -4.171      0.000      -0.907      -0.327
==============================================================================
Omnibus:                       24.832   Durbin-Watson:                   1.790
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.637
Skew:                          -0.618   Prob(JB):                     9.97e-07
Kurtosis:                       3.223   Cond. No.                         18.0
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iii. UEVT (436143) vs DMS (578209)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.126
Model:                                  OLS   Adj. R-squared:                  0.111
Method:                       Least Squares   F-statistic:                     9.188
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           6.36e-07
Time:                              11:10:52   Log-Likelihood:                -21.154
No. Observations:                       245   AIC:                             52.31
Df Residuals:                           240   BIC:                             69.81
Df Model:                                 4
Covariance Type:                        HC1
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                    0.6575      0.082      7.984      0.000       0.496       0.819
author                                  -0.0117      0.034     -0.342      0.732      -0.079       0.055
pretest_avg_problem_accuracy             0.1684      0.076      2.214      0.027       0.019       0.317
student_std_attempted                   -0.0555      0.245     -0.226      0.821      -0.536       0.425
student_std_attempted_before_support    -0.5778      0.224     -2.579      0.010      -1.017      -0.139
==============================================================================
Omnibus:                       24.999   Durbin-Watson:                   1.912
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.418
Skew:                          -0.815   Prob(JB):                     4.09e-07
Kurtosis:                       3.473   Cond. No.                         19.8
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iv. UEVT (436143) vs EGS (579215)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.187
Model:                                  OLS   Adj. R-squared:                  0.167
Method:                       Least Squares   F-statistic:                     9.254
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           9.43e-07
Time:                              11:10:53   Log-Likelihood:                -13.912
No. Observations:                       165   AIC:                             37.82
Df Residuals:                           160   BIC:                             53.35
Df Model:                                 4
Covariance Type:                        HC1
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                    0.7097      0.082      8.663      0.000       0.549       0.870
author                                   0.0351      0.041      0.854      0.393      -0.045       0.116
pretest_avg_problem_accuracy             0.1553      0.067      2.315      0.021       0.024       0.287
student_std_attempted                   -0.2578      0.304     -0.847      0.397      -0.854       0.338
student_std_attempted_before_support    -0.7015      0.221     -3.174      0.002      -1.135      -0.268
==============================================================================
Omnibus:                        5.634   Durbin-Watson:                   1.862
Prob(Omnibus):                  0.060   Jarque-Bera (JB):                5.605
Skew:                          -0.413   Prob(JB):                       0.0607
Kurtosis:                       2.637   Cond. No.                         16.2
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

v. UEVT (436143) vs BTS (579216)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.264
Model:                                  OLS   Adj. R-squared:                  0.239
Method:                       Least Squares   F-statistic:                     12.25
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.19e-08
Time:                              11:10:54   Log-Likelihood:                -15.208
No. Observations:                       125   AIC:                             40.42
Df Residuals:                           120   BIC:                             54.56
Df Model:                                 4
Covariance Type:                        HC1
==========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                   0.8137      0.110      7.408      0.000       0.598       1.029
author                                 -0.0404      0.049     -0.823      0.411      -0.137       0.056
pretest_avg_problem_accuracy            0.1993      0.090      2.207      0.027       0.022       0.376
student_std_attempted                  -0.6547      0.344     -1.901      0.057      -1.330       0.020
student_std_attempted_before_support   -0.8725      0.221     -3.943      0.000      -1.306      -0.439
==============================================================================
Omnibus:                        2.896   Durbin-Watson:                   1.584
Prob(Omnibus):                  0.235   Jarque-Bera (JB):                2.791
Skew:                          -0.363   Prob(JB):                        0.248
Kurtosis:                       2.910   Cond. No.                         15.9
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

vi. UEVT (436143) vs BGS (579218)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.154
Model:                                  OLS   Adj. R-squared:                  0.127
Method:                       Least Squares   F-statistic:                     6.636
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           7.14e-05
Time:                              11:10:55   Log-Likelihood:                -15.643
No. Observations:                       129   AIC:                             41.29
Df Residuals:                           124   BIC:                             55.59
Df Model:                                 4
Covariance Type:                        HC1
==========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                   0.5986      0.120      4.991      0.000       0.364       0.834
author                                  0.0548      0.051      1.069      0.285      -0.046       0.155
pretest_avg_problem_accuracy            0.1612      0.089      1.806      0.071      -0.014       0.336
student_std_attempted                   0.1326      0.258      0.513      0.608      -0.373       0.639
student_std_attempted_before_support   -0.7807      0.249     -3.130      0.002      -1.270      -0.292
==============================================================================
Omnibus:                        8.681   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.013   Jarque-Bera (JB):                9.193
Skew:                          -0.652   Prob(JB):                       0.0101
Kurtosis:                       2.905   Cond. No.                         15.7
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

vii. BCT (436919) vs LQS (579217)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.243
Model:                                  OLS   Adj. R-squared:                  0.237
Method:                       Least Squares   F-statistic:                     44.66
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.61e-32
Time:                              11:10:58   Log-Likelihood:                -41.756
No. Observations:                       533   AIC:                             93.51
Df Residuals:                           528   BIC:                             114.9
Df Model:                                 4
Covariance Type:                        HC1
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     0.6493      0.050     12.862      0.000       0.550       0.748
author                                    0.0367      0.023      1.617      0.106      -0.008       0.081
pretest_avg_problem_accuracy              0.2271      0.050      4.509      0.000       0.128       0.326
student_std_attempted                    -0.0752      0.147     -0.510      0.610      -0.364       0.214
student_std_attempted_before_support     -0.9141      0.111     -8.259      0.000      -1.131      -0.697
==============================================================================
Omnibus:                       17.976   Durbin-Watson:                   1.621
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.879
Skew:                          -0.452   Prob(JB):                     7.95e-05
Kurtosis:                       3.184   Cond. No.                         17.1
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

viii. BCT (436919) vs BGS (579218)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.264
Model:                                  OLS   Adj. R-squared:                  0.252
Method:                       Least Squares   F-statistic:                     20.48
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           1.50e-14
Time:                              11:11:00   Log-Likelihood:                -16.430
No. Observations:                       247   AIC:                             42.86
Df Residuals:                           242   BIC:                             60.41
Df Model:                                 4
Covariance Type:                        HC1
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     0.5730      0.086      6.636      0.000       0.404       0.742
author                                   -0.0042      0.034     -0.126      0.899      -0.070       0.061
pretest_avg_problem_accuracy              0.3207      0.081      3.983      0.000       0.163       0.479
student_std_attempted                    -0.1075      0.203     -0.529      0.597      -0.505       0.291
student_std_attempted_before_support     -0.7055      0.177     -3.995      0.000      -1.052      -0.359
==============================================================================
Omnibus:                        2.144   Durbin-Watson:                   1.857
Prob(Omnibus):                  0.342   Jarque-Bera (JB):                1.737
Skew:                          -0.037   Prob(JB):                        0.420
Kurtosis:                       2.596   Cond. No.                         14.5
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

ix. BCT (436919) vs NQS (579220)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.283
Model:                                  OLS   Adj. R-squared:                  0.275
Method:                       Least Squares   F-statistic:                     39.57
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.04e-27
Time:                              11:11:02   Log-Likelihood:                -19.211
No. Observations:                       361   AIC:                             48.42
Df Residuals:                           356   BIC:                             67.87
Df Model:                                 4
Covariance Type:                        HC1
======================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------
const                                   0.7095      0.066     10.755      0.000       0.580       0.839
author                                 -0.0159      0.027     -0.580      0.562      -0.069       0.038
pretest_avg_problem_accuracy            0.1927      0.063      3.051      0.002       0.069       0.317
student_std_attempted                  -0.1490      0.151     -0.983      0.325      -0.446       0.148
student_std_attempted_before_support   -1.0157      0.134     -7.586      0.000      -1.278      -0.753
==============================================================================
Omnibus:                        6.102   Durbin-Watson:                   1.687
Prob(Omnibus):                  0.047   Jarque-Bera (JB):                6.244
Skew:                          -0.308   Prob(JB):                       0.0441
Kurtosis:                       2.810   Cond. No.                         15.5
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

The next section shows the five different regression results for Study 2 for models with main effects. Table 3 summarized a few key results from the below five regressions, and that none of them allow us to reliably say one teacher is better than another. Please note that the following indices use the same author pairs as in Study 1:

Study 1	Study 2
i	i
ii	ii
iii	iii
vii	iv
ix	v

i. UEVT (436143) vs OTS (578207)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.282
Model:                                  OLS   Adj. R-squared:                  0.252
Method:                       Least Squares   F-statistic:                     12.55
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.98e-08
Time:                              11:13:47   Log-Likelihood:                -11.518
No. Observations:                       101   AIC:                             33.04
Df Residuals:                            96   BIC:                             46.11
Df Model:                                 4
Covariance Type:                        HC1
========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                   0.5481      0.129      4.263      0.000       0.296       0.800
author                                  0.0025      0.057      0.044      0.965      -0.108       0.113
pretest_avg_problem_accuracy            0.3324      0.117      2.843      0.004       0.103       0.562
student_std_attempted                  -0.2047      0.378     -0.541      0.588      -0.946       0.537
student_std_attempted_before_support   -1.0457      0.341     -3.063      0.002      -1.715      -0.377
==============================================================================
Omnibus:                        3.688   Durbin-Watson:                   1.370
Prob(Omnibus):                  0.158   Jarque-Bera (JB):                3.673
Skew:                          -0.455   Prob(JB):                        0.159
Kurtosis:                       2.786   Cond. No.                         19.9
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

ii. UEVT (436143) vs JXS (578208)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.152
Model:                                  OLS   Adj. R-squared:                  0.121
Method:                       Least Squares   F-statistic:                     4.641
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):            0.00169
Time:                              11:13:48   Log-Likelihood:                -18.712
No. Observations:                       114   AIC:                             47.42
Df Residuals:                           109   BIC:                             61.11
Df Model:                                 4
Covariance Type:                        HC1
========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                   0.7821      0.127      6.181      0.000       0.534       1.030
author                                 -0.0242      0.055     -0.438      0.661      -0.132       0.084
pretest_avg_problem_accuracy            0.0868      0.119      0.727      0.467      -0.147       0.321
student_std_attempted                  -0.2552      0.417     -0.613      0.540      -1.072       0.561
student_std_attempted_before_support   -0.9208      0.340     -2.709      0.007      -1.587      -0.255
==============================================================================
Omnibus:                        6.179   Durbin-Watson:                   1.763
Prob(Omnibus):                  0.046   Jarque-Bera (JB):                6.377
Skew:                          -0.561   Prob(JB):                       0.0412
Kurtosis:                       2.707   Cond. No.                         21.5
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iii. UEVT (436143) vs DMS (578209)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.086
Model:                                  OLS   Adj. R-squared:                  0.055
Method:                       Least Squares   F-statistic:                     2.728
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):             0.0325
Time:                              11:13:49   Log-Likelihood:                -17.222
No. Observations:                       123   AIC:                             44.44
Df Residuals:                           118   BIC:                             58.50
Df Model:                                 4
Covariance Type:                        HC1
========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                   0.6103      0.101      6.055      0.000       0.413       0.808
author                                  0.0146      0.051      0.286      0.775      -0.085       0.115
pretest_avg_problem_accuracy            0.1511      0.104      1.460      0.144      -0.052       0.354
student_std_attempted                  -0.1161      0.326     -0.356      0.722      -0.756       0.523
student_std_attempted_before_support   -0.5606      0.267     -2.096      0.036      -1.085      -0.036
==============================================================================
Omnibus:                        9.901   Durbin-Watson:                   1.559
Prob(Omnibus):                  0.007   Jarque-Bera (JB):                8.620
Skew:                          -0.565   Prob(JB):                       0.0134
Kurtosis:                       2.365   Cond. No.                         18.5
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iv. BCT (436919) vs LQS (579217)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.163
Model:                                  OLS   Adj. R-squared:                  0.154
Method:                       Least Squares   F-statistic:                     18.86
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           4.37e-14
Time:                              11:13:51   Log-Likelihood:                -35.050
No. Observations:                       364   AIC:                             80.10
Df Residuals:                           359   BIC:                             99.58
Df Model:                                 4
Covariance Type:                        HC1
========================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                   0.5671      0.059      9.594      0.000       0.451       0.683
author                                  0.0268      0.028      0.955      0.340      -0.028       0.082
pretest_avg_problem_accuracy            0.2262      0.054      4.209      0.000       0.121       0.332
student_std_attempted                  -0.3591      0.225     -1.595      0.111      -0.800       0.082
student_std_attempted_before_support   -0.4751      0.127     -3.751      0.000      -0.723      -0.227
==============================================================================
Omnibus:                        4.112   Durbin-Watson:                   1.820
Prob(Omnibus):                  0.128   Jarque-Bera (JB):                4.049
Skew:                          -0.220   Prob(JB):                        0.132
Kurtosis:                       2.729   Cond. No.                         18.4
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

v. BCT (436919) vs NQS (579220)

```
                            OLS Regression Results
=====================================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.176
Model:                                  OLS   Adj. R-squared:                  0.162
Method:                       Least Squares   F-statistic:                     14.76
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           8.80e-11
Time:                              11:13:52   Log-Likelihood:                -26.909
No. Observations:                       242   AIC:                             63.82
Df Residuals:                           237   BIC:                             81.26
Df Model:                                 4
Covariance Type:                        HC1
=========================================================================================================
                                      coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                               0.5335      0.070      7.592      0.000       0.396       0.671
author                             -0.0108      0.035     -0.304      0.761      -0.080       0.059
pretest_avg_problem_accuracy        0.2216      0.062      3.554      0.000       0.099       0.344
student_std_attempted               0.0535      0.264      0.203      0.839      -0.464       0.571
student_std_attempted_before_support   -0.6150      0.160     -3.856      0.000      -0.928      -0.302
==============================================================================
Omnibus:                        3.740   Durbin-Watson:                   1.657
Prob(Omnibus):                  0.154   Jarque-Bera (JB):                3.678
Skew:                          -0.258   Prob(JB):                        0.159
Kurtosis:                       2.687   Cond. No.                         20.2
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

Appendix B: RQ2 Results for Study 1 and 2

Recall that RQ2 is "When comparing authors, are there reliable differences based upon demographics?" To answer this question, we computed regressions that included both demographics and interaction terms. Not too surprisingly, the demographics features helped predict posttest scores, but the real question is about the interaction terms. If there are reliable interactions between authors and any of the demographics features, that would suggest that some group of students learn better with one of the teachers versus the other teacher. We have found little evidence to suggest any such reliable heterogeneous treatment effects.

Note that the number of observations appears lower. This is due to the fact that we are missing demographic information for some students. Please note that since this model has an intercept (labeled as const) representing an urban, high knowledge, female.

This section shows the nine different regression results for Study 1 for models with main effects. Table 2 summarized the results in the ‘Demographic’ column.

i. UEVT (436143) vs OTS (578207)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.350
Model:                                  OLS   Adj. R-squared:                  0.204
Method:                       Least Squares   F-statistic:                     5.167
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.61e-06
Time:                              11:11:04   Log-Likelihood:                 8.0020
No. Observations:                        77   AIC:                             14.00
Df Residuals:                            62   BIC:                             49.15
Df Model:                                14
Covariance Type:                        HC1
===================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                                   0.9370      0.236      3.971      0.000       0.475       1.400
author                                 -0.1161      0.173     -0.671      0.502      -0.455       0.223
pretest_avg_problem_accuracy           -0.0612      0.152     -0.403      0.687      -0.359       0.236
student_std_attempted                  -0.5279      0.477     -1.106      0.269      -1.463       0.407
student_std_attempted_before_support    0.4635      0.557      0.832      0.405      -0.628       1.555
gender                                  0.0041      0.067      0.062      0.951      -0.127       0.135
low_knowledge                          -0.4663      0.186     -2.502      0.012      -0.832      -0.101
mid_knowledge                          -0.2733      0.116     -2.352      0.019      -0.501      -0.046
rural                                   0.0392      0.143      0.274      0.784      -0.240       0.319
suburban                               -0.0055      0.104     -0.053      0.958      -0.208       0.198
author:rural                            0.1192      0.163      0.729      0.466      -0.201       0.440
author:suburban                        -0.0776      0.187     -0.416      0.677      -0.443       0.288
author:low_knowledge                    0.1873      0.173      1.081      0.280      -0.152       0.527
author:mid_knowledge                    0.1465      0.136      1.081      0.280      -0.119       0.412
author:gender                          -0.0253      0.099     -0.256      0.798      -0.219       0.168
==============================================================================
Omnibus:                       16.509   Durbin-Watson:                   1.755
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.442
Skew:                          -0.990   Prob(JB):                     3.64e-05
Kurtosis:                       4.566   Cond. No.                         36.7
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

ii. UEVT (436143) vs JXS (578208)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.345
Model:                                  OLS   Adj. R-squared:                  0.287
Method:                       Least Squares   F-statistic:                     7.048
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           3.60e-11
Time:                              11:11:06   Log-Likelihood:                 19.982
No. Observations:                       173   AIC:                            -9.964
Df Residuals:                           158   BIC:                             37.34
Df Model:                                14
Covariance Type:                        HC1
===================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                                   0.7387      0.116      6.350      0.000       0.511       0.967
author                                 -0.0986      0.083     -1.190      0.234      -0.261       0.064
pretest_avg_problem_accuracy            0.2921      0.087      3.359      0.001       0.122       0.462
student_std_attempted                  -0.2483      0.322     -0.772      0.440      -0.878       0.382
student_std_attempted_before_support   -0.0799      0.322     -0.249      0.804      -0.710       0.550
gender                                 -0.0841      0.059     -1.421      0.155      -0.200       0.032
low_knowledge                          -0.1960      0.100     -1.963      0.050      -0.392      -0.000
mid_knowledge                          -0.0544      0.055     -0.991      0.322      -0.162       0.053
rural                                  -0.0719      0.068     -1.064      0.287      -0.204       0.060
suburban                               -0.1258      0.065     -1.926      0.054      -0.254       0.002
author:rural                            0.1416      0.090      1.580      0.114      -0.034       0.317
author:suburban                         0.1074      0.097      1.109      0.267      -0.082       0.297
author:low_knowledge                   -0.0368      0.103     -0.356      0.722      -0.239       0.166
author:mid_knowledge                   -0.0689      0.069     -1.004      0.315      -0.203       0.066
author:gender                           0.0740      0.073      1.020      0.308      -0.068       0.216
==============================================================================
Omnibus:                        7.734   Durbin-Watson:                   1.724
Prob(Omnibus):                  0.021   Jarque-Bera (JB):                7.686
Skew:                          -0.427   Prob(JB):                       0.0214
Kurtosis:                       3.581   Cond. No.                         29.4
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iii. UEVT (436143) vs DMS (578209)

```
                             OLS Regression Results
==================================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.208
Model:                                  OLS   Adj. R-squared:                  0.058
Method:                       Least Squares   F-statistic:                     4.869
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.86e-06
Time:                              11:11:07   Log-Likelihood:                -6.6861
No. Observations:                        89   AIC:                             43.37
Df Residuals:                            74   BIC:                             80.70
Df Model:                                14
Covariance Type:                        HC1
====================================================================================================
                                        coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                                 0.9761      0.224      4.365      0.000       0.538       1.414
author                                0.0090      0.196      0.046      0.963      -0.376       0.394
pretest_avg_problem_accuracy          0.0483      0.122      0.395      0.693      -0.192       0.288
student_std_attempted                -0.1815      0.584     -0.311      0.756      -1.327       0.964
student_std_attempted_before_support -0.7592      0.534     -1.422      0.155      -1.805       0.287
gender                               -0.0926      0.119     -0.776      0.438      -0.326       0.141
low_knowledge                         0.0573      0.165      0.347      0.728      -0.266       0.380
mid_knowledge                        -0.0282      0.160     -0.176      0.860      -0.342       0.286
rural                                -0.2251      0.108     -2.079      0.038      -0.437      -0.013
suburban                             -0.1215      0.152     -0.798      0.425      -0.420       0.177
author:rural                          0.1876      0.130      1.448      0.148      -0.066       0.442
author:suburban                       0.1127      0.183      0.615      0.538      -0.246       0.472
author:low_knowledge                 -0.2244      0.183     -1.227      0.220      -0.583       0.134
author:mid_knowledge                 -0.1868      0.170     -1.099      0.272      -0.520       0.146
author:gender                         0.1002      0.138      0.725      0.469      -0.171       0.371
==============================================================================
Omnibus:                        5.939   Durbin-Watson:                   1.816
Prob(Omnibus):                  0.051   Jarque-Bera (JB):                5.635
Skew:                          -0.614   Prob(JB):                       0.0598
Kurtosis:                       3.105   Cond. No.                         35.6
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iv. UEVT (436143) vs EGS (579215)

```
                             OLS Regression Results
==================================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.637
Model:                                  OLS   Adj. R-squared:                  0.395
Method:                       Least Squares   F-statistic:                     50.56
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           5.70e-13
Time:                              11:11:08   Log-Likelihood:                 17.059
No. Observations:                        36   AIC:                            -4.119
Df Residuals:                            21   BIC:                             19.63
Df Model:                                14
Covariance Type:                        HC1
====================================================================================================
                                        coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                                 1.2871      0.161      7.997      0.000       0.972       1.603
author                               -0.2819      0.135     -2.088      0.037      -0.547      -0.017
pretest_avg_problem_accuracy         -0.0387      0.128     -0.302      0.763      -0.290       0.212
student_std_attempted                -0.4885      0.661     -0.739      0.460      -1.784       0.807
student_std_attempted_before_support -1.3071      0.515     -2.537      0.011      -2.317      -0.297
gender                                0.3059      0.061      4.985      0.000       0.186       0.426
low_knowledge                        -0.0397      0.135     -0.294      0.769      -0.304       0.225
mid_knowledge                        -0.3256      0.145     -2.250      0.024      -0.609      -0.042
rural                                -0.7571      0.097     -7.832      0.000      -0.947      -0.568
suburban                             -0.3833      0.114     -3.352      0.001      -0.607      -0.159
author:rural                          0.6198      0.222      2.791      0.005       0.185       1.055
author:suburban                       0.3474      0.188      1.848      0.065      -0.021       0.716
author:low_knowledge                  0.1376      0.140      0.985      0.325      -0.136       0.411
author:mid_knowledge                  0.2206      0.266      0.829      0.407      -0.301       0.742
author:gender                        -0.1164      0.177     -0.658      0.510      -0.463       0.230
==============================================================================
Omnibus:                        0.637   Durbin-Watson:                   1.548
Prob(Omnibus):                  0.727   Jarque-Bera (JB):                0.675
Skew:                           0.018   Prob(JB):                        0.714
Kurtosis:                       2.330   Cond. No.                         35.9
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

v. UEVT (436143) vs BTS (579216)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.586
Model:                                  OLS   Adj. R-squared:                  0.473
Method:                       Least Squares   F-statistic:                     11.15
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           4.85e-11
Time:                              11:11:09   Log-Likelihood:                 25.954
No. Observations:                        66   AIC:                            -21.91
Df Residuals:                            51   BIC:                             10.94
Df Model:                                14
Covariance Type:                        HC1
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     0.9278      0.180      5.159      0.000       0.575       1.280
author                                   -0.1370      0.143     -0.960      0.337      -0.417       0.143
pretest_avg_problem_accuracy              0.2151      0.111      1.933      0.053      -0.003       0.433
student_std_attempted                    -1.0641      0.256     -4.151      0.000      -1.567      -0.562
student_std_attempted_before_support     -0.6074      0.486     -1.250      0.211      -1.560       0.345
gender                                    0.0439      0.114      0.385      0.700      -0.180       0.267
low_knowledge                            -0.1542      0.113     -1.365      0.172      -0.376       0.067
mid_knowledge                             0.0797      0.103      0.775      0.438      -0.122       0.281
rural                                    -0.0952      0.103     -0.921      0.357      -0.298       0.107
suburban                                 -0.0435      0.137     -0.318      0.750      -0.312       0.225
author:rural                              0.1542      0.120      1.284      0.199      -0.081       0.389
author:suburban                           0.0390      0.148      0.264      0.792      -0.251       0.329
author:low_knowledge                      0.1404      0.148      0.946      0.344      -0.151       0.431
author:mid_knowledge                     -0.0954      0.125     -0.763      0.445      -0.341       0.150
author:gender                            -0.1075      0.125     -0.862      0.389      -0.352       0.137
==============================================================================
Omnibus:                        0.117   Durbin-Watson:                   1.588
Prob(Omnibus):                  0.943   Jarque-Bera (JB):                0.280
Skew:                           0.076   Prob(JB):                        0.869
Kurtosis:                       2.720   Cond. No.                         31.4
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

vi. UEVT (436143) vs BGS (579218)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.660
Model:                                  OLS   Adj. R-squared:                  0.409
Method:                       Least Squares   F-statistic:                     4.257
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):            0.00203
Time:                              11:11:10   Log-Likelihood:                 6.3269
No. Observations:                        34   AIC:                             17.35
Df Residuals:                            19   BIC:                             40.24
Df Model:                                14
Covariance Type:                        HC1
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     1.1220      0.295      3.798      0.000       0.543       1.701
author                                   -0.2801      0.384     -0.730      0.465      -1.032       0.472
pretest_avg_problem_accuracy              0.1331      0.225      0.592      0.554      -0.307       0.573
student_std_attempted                    -0.7434      0.773     -0.962      0.336      -2.258       0.772
student_std_attempted_before_support     -3.0036      0.846     -3.548      0.000      -4.663      -1.344
gender                                    0.1545      0.209      0.740      0.459      -0.255       0.564
low_knowledge                             0.7395      0.271      2.733      0.006       0.209       1.270
mid_knowledge                             0.6241      0.176      3.539      0.000       0.278       0.970
rural                                    -0.5736      0.206     -2.787      0.005      -0.977      -0.170
suburban                                 -0.6986      0.143     -4.900      0.000      -0.978      -0.419
author:rural                              0.2008      0.400      0.502      0.616      -0.583       0.984
author:suburban                           0.1304      0.377      0.346      0.729      -0.608       0.869
author:low_knowledge                      0.1224      0.413      0.296      0.767      -0.687       0.932
author:mid_knowledge                      0.1970      0.409      0.481      0.630      -0.605       0.999
author:gender                            -0.0487      0.257     -0.190      0.850      -0.552       0.455
==============================================================================
Omnibus:                        3.050   Durbin-Watson:                   1.409
Prob(Omnibus):                  0.218   Jarque-Bera (JB):                2.741
Skew:                          -0.625   Prob(JB):                        0.254
Kurtosis:                       2.391   Cond. No.                         38.7
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

vii. BCT (436919) vs LQS (579217)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.380
Model:                                  OLS   Adj. R-squared:                  0.352
Method:                       Least Squares   F-statistic:                     13.04
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           6.10e-24
Time:                              11:11:13   Log-Likelihood:                 30.942
No. Observations:                       315   AIC:                            -31.88
Df Residuals:                           300   BIC:                             24.40
Df Model:                                14
Covariance Type:                        HC1
==================================================================================================
                                        coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                                 0.6917      0.080      8.643      0.000       0.535       0.849
author                                0.0300      0.051      0.583      0.560      -0.071       0.131
pretest_avg_problem_accuracy          0.2610      0.066      3.951      0.000       0.132       0.390
student_std_attempted                -0.2270      0.258     -0.881      0.378      -0.732       0.278
student_std_attempted_before_support -0.7122      0.237     -3.004      0.003      -1.177      -0.248
gender                                0.0471      0.040      1.164      0.244      -0.032       0.126
low_knowledge                        -0.1036      0.069     -1.509      0.131      -0.238       0.031
mid_knowledge                        -0.0418      0.038     -1.100      0.271      -0.116       0.033
rural                                -0.1366      0.057     -2.404      0.016      -0.248      -0.025
suburban                             -0.0118      0.043     -0.274      0.784      -0.096       0.073
author:rural                         -0.0464      0.084     -0.555      0.579      -0.210       0.117
author:suburban                       0.0559      0.054      1.037      0.300      -0.050       0.161
author:low_knowledge                  0.0043      0.067      0.065      0.948      -0.127       0.135
author:mid_knowledge                  0.0372      0.051      0.725      0.468      -0.063       0.138
author:gender                         0.0094      0.054      0.176      0.861      -0.096       0.115
==============================================================================
Omnibus:                       19.724   Durbin-Watson:                   1.269
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.360
Skew:                          -0.564   Prob(JB):                     1.39e-05
Kurtosis:                       3.658   Cond. No.                         34.2
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

viii. BCT (436919) vs BGS (579218)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.306
Model:                                  OLS   Adj. R-squared:                  0.243
Method:                       Least Squares   F-statistic:                     6.662
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           1.81e-10
Time:                              11:11:15   Log-Likelihood:                 23.658
No. Observations:                       169   AIC:                            -17.32
Df Residuals:                           154   BIC:                             29.63
Df Model:                                14
Covariance Type:                        HC1
==================================================================================================
                                        coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                                 0.6023      0.129      4.665      0.000       0.349       0.855
author                                0.0158      0.083      0.192      0.848      -0.146       0.178
pretest_avg_problem_accuracy          0.3446      0.125      2.751      0.006       0.099       0.590
student_std_attempted                 0.2201      0.382      0.577      0.564      -0.528       0.968
student_std_attempted_before_support -0.6535      0.287     -2.280      0.023      -1.215      -0.092
gender                               -0.0298      0.042     -0.715      0.474      -0.111       0.052
low_knowledge                        -0.0124      0.079     -0.158      0.875      -0.166       0.142
mid_knowledge                        -0.0233      0.061     -0.385      0.700      -0.142       0.096
rural                                 0.0223      0.051      0.436      0.663      -0.078       0.122
suburban                             -0.0771      0.064     -1.212      0.226      -0.202       0.048
author:rural                          0.0080      0.077      0.104      0.917      -0.142       0.158
author:suburban                       0.0776      0.102      0.759      0.448      -0.123       0.278
author:low_knowledge                 -0.0143      0.088     -0.162      0.871      -0.187       0.158
author:mid_knowledge                 -0.0244      0.080     -0.303      0.762      -0.182       0.133
author:gender                         0.0427      0.070      0.608      0.543      -0.095       0.180
==============================================================================
Omnibus:                        9.765   Durbin-Watson:                   1.489
Prob(Omnibus):                  0.008   Jarque-Bera (JB):                9.851
Skew:                          -0.524   Prob(JB):                      0.00726
Kurtosis:                       3.549   Cond. No.                         34.0
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

ix. BCT (436919) vs NQS (579220)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.312
Model:                                  OLS   Adj. R-squared:                  0.267
Method:                       Least Squares   F-statistic:                     6.936
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           1.11e-11
Time:                              11:11:17   Log-Likelihood:                 36.700
No. Observations:                       228   AIC:                            -43.40
Df Residuals:                           213   BIC:                             8.041
Df Model:                                14
Covariance Type:                        HC1
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                    0.8644      0.078     11.127      0.000       0.712       1.017
author                                   0.0820      0.083      0.992      0.321      -0.080       0.244
pretest_avg_problem_accuracy             0.1004      0.075      1.347      0.178      -0.046       0.247
student_std_attempted                   -0.4900      0.290     -1.690      0.091      -1.058       0.078
student_std_attempted_before_support    -0.4633      0.266     -1.739      0.082      -0.985       0.059
gender                                  -0.1467      0.039     -3.720      0.000      -0.224      -0.069
low_knowledge                           -0.1204      0.074     -1.626      0.104      -0.265       0.025
mid_knowledge                            0.0298      0.054      0.547      0.584      -0.077       0.136
rural                                    0.0457      0.051      0.898      0.369      -0.054       0.145
suburban                                 0.0405      0.052      0.780      0.435      -0.061       0.142
author:rural                            -0.1999      0.087     -2.302      0.021      -0.370      -0.030
author:suburban                         -0.1360      0.069     -1.967      0.049      -0.271      -0.000
author:low_knowledge                    -0.0408      0.086     -0.473      0.636      -0.210       0.128
author:mid_knowledge                    -0.1030      0.078     -1.322      0.186      -0.256       0.050
author:gender                            0.1146      0.056      2.052      0.040       0.005       0.224
==============================================================================
Omnibus:                        6.844   Durbin-Watson:                   1.243
Prob(Omnibus):                  0.033   Jarque-Bera (JB):                6.569
Skew:                          -0.380   Prob(JB):                       0.0375
Kurtosis:                       3.340   Cond. No.                         36.7
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

The next section shows the five different regression results for Study 2 for models with main effects. Table 3 summarized the results in the ‘Demographic’ column.

i. UEVT (436143) vs OTS (578207)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.636
Model:                                  OLS   Adj. R-squared:                  0.455
Method:                       Least Squares   F-statistic:                     11.89
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.69e-08
Time:                              11:13:53   Log-Likelihood:                 4.6567
No. Observations:                        43   AIC:                             20.69
Df Residuals:                            28   BIC:                             47.10
Df Model:                                14
Covariance Type:                        HC1
======================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------
const                                   0.7183      0.218      3.295      0.001       0.291       1.146
author                                 -0.6684      0.243     -2.750      0.006      -1.145      -0.192
pretest_avg_problem_accuracy            0.6202      0.199      3.109      0.002       0.229       1.011
student_std_attempted                  -0.8714      0.619     -1.407      0.159      -2.085       0.342
student_std_attempted_before_support   -1.6548      0.713     -2.320      0.020      -3.053      -0.257
gender                                  0.0186      0.087      0.215      0.830      -0.151       0.188
low_knowledge                           0.0751      0.276      0.272      0.785      -0.466       0.616
mid_knowledge                          -0.0331      0.140     -0.236      0.813      -0.307       0.241
rural                                  -0.2351      0.146     -1.611      0.107      -0.521       0.051
suburban                               -0.0484      0.127     -0.382      0.703      -0.297       0.200
author:rural                            0.1852      0.283      0.653      0.514      -0.370       0.741
author:suburban                         0.2762      0.227      1.219      0.223      -0.168       0.720
author:low_knowledge                    0.2511      0.316      0.794      0.427      -0.369       0.871
author:mid_knowledge                    0.5618      0.320      1.757      0.079      -0.065       1.188
author:gender                           0.0067      0.176      0.038      0.970      -0.339       0.353
==============================================================================
Omnibus:                        0.027   Durbin-Watson:                   1.463
Prob(Omnibus):                  0.986   Jarque-Bera (JB):                0.121
Skew:                           0.053   Prob(JB):                        0.941
Kurtosis:                       2.763   Cond. No.                         31.8
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

ii. UEVT (436143) vs JXS (578208)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.440
Model:                                  OLS   Adj. R-squared:                  0.228
Method:                       Least Squares   F-statistic:                     4.583
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           0.000101
Time:                              11:13:55   Log-Likelihood:                 1.2041
No. Observations:                        52   AIC:                             27.59
Df Residuals:                            37   BIC:                             56.86
Df Model:                                14
Covariance Type:                        HC1
======================================================================================================
                                          coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------
const                                   1.2885      0.211      6.094      0.000       0.874       1.703
author                                 -0.3442      0.179     -1.923      0.054      -0.695       0.007
pretest_avg_problem_accuracy           -0.0535      0.158     -0.339      0.735      -0.363       0.256
student_std_attempted                   0.1818      0.692      0.263      0.793      -1.174       1.537
student_std_attempted_before_support   -1.8668      0.613     -3.047      0.002      -3.068      -0.666
gender                                  0.0251      0.133      0.188      0.851      -0.236       0.286
low_knowledge                          -0.0406      0.171     -0.238      0.812      -0.375       0.294
mid_knowledge                          -0.3589      0.140     -2.558      0.011      -0.634      -0.084
rural                                  -0.3048      0.140     -2.183      0.029      -0.579      -0.031
suburban                               -0.0520      0.148     -0.351      0.726      -0.342       0.239
author:rural                            0.1809      0.190      0.952      0.341      -0.191       0.553
author:suburban                         0.0345      0.238      0.145      0.885      -0.432       0.501
author:low_knowledge                   -0.0803      0.215     -0.374      0.708      -0.501       0.341
author:mid_knowledge                    0.4632      0.212      2.184      0.029       0.048       0.879
author:gender                           0.0663      0.167      0.396      0.692      -0.262       0.395
==============================================================================
Omnibus:                        1.414   Durbin-Watson:                   2.155
Prob(Omnibus):                  0.493   Jarque-Bera (JB):                1.419
Skew:                          -0.331   Prob(JB):                        0.492
Kurtosis:                       2.535   Cond. No.                         31.5
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iii. UEVT (436143) vs DMS (578209)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.358
Model:                                  OLS   Adj. R-squared:                  0.167
Method:                       Least Squares   F-statistic:                     6.386
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           6.43e-07
Time:                              11:13:56   Log-Likelihood:                 5.3689
No. Observations:                        62   AIC:                             19.26
Df Residuals:                            47   BIC:                             51.17
Df Model:                                14
Covariance Type:                        HC1
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     1.2101      0.261      4.638      0.000       0.699       1.721
author                                   -0.2114      0.256     -0.825      0.409      -0.713       0.291
pretest_avg_problem_accuracy              0.0801      0.156      0.515      0.607      -0.225       0.385
student_std_attempted                    -0.5430      0.320     -1.697      0.090      -1.170       0.084
student_std_attempted_before_support     -0.3587      0.392     -0.916      0.360      -1.127       0.409
gender                                   -0.1147      0.133     -0.865      0.387      -0.375       0.145
low_knowledge                            -0.3938      0.246     -1.601      0.109      -0.876       0.088
mid_knowledge                            -0.3569      0.109     -3.266      0.001      -0.571      -0.143
rural                                    -0.3312      0.196     -1.691      0.091      -0.715       0.053
suburban                                 -0.1502      0.196     -0.765      0.444      -0.535       0.235
author:rural                             -0.0059      0.228     -0.026      0.979      -0.453       0.441
author:suburban                          -0.1626      0.219     -0.742      0.458      -0.592       0.267
author:low_knowledge                      0.2229      0.210      1.061      0.289      -0.189       0.635
author:mid_knowledge                      0.1360      0.133      1.026      0.305      -0.124       0.396
author:gender                             0.2968      0.160      1.851      0.064      -0.017       0.611
==============================================================================
Omnibus:                        3.646   Durbin-Watson:                   1.083
Prob(Omnibus):                  0.162   Jarque-Bera (JB):                3.004
Skew:                          -0.534   Prob(JB):                        0.223
Kurtosis:                       3.152   Cond. No.                         36.5
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

iv. BCT (436919) vs LQS (579217)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.321
Model:                                  OLS   Adj. R-squared:                  0.280
Method:                       Least Squares   F-statistic:                     11.15
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           2.32e-19
Time:                              11:13:57   Log-Likelihood:                 15.757
No. Observations:                       246   AIC:                            -1.514
Df Residuals:                           231   BIC:                             51.07
Df Model:                                14
Covariance Type:                        HC1
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     0.9999      0.112      8.920      0.000       0.780       1.220
author                                   -0.2430      0.093     -2.616      0.009      -0.425      -0.061
pretest_avg_problem_accuracy              0.0612      0.061      1.008      0.313      -0.058       0.180
student_std_attempted                    -1.0015      0.333     -3.006      0.003      -1.654      -0.349
student_std_attempted_before_support     -0.1197      0.263     -0.455      0.649      -0.635       0.396
gender                                    0.0781      0.056      1.399      0.162      -0.031       0.187
low_knowledge                            -0.1600      0.085     -1.885      0.059      -0.326       0.006
mid_knowledge                            -0.0570      0.069     -0.824      0.410      -0.193       0.079
rural                                    -0.4125      0.059     -7.018      0.000      -0.528      -0.297
suburban                                 -0.3296      0.053     -6.254      0.000      -0.433      -0.226
author:rural                              0.3642      0.070      5.169      0.000       0.226       0.502
author:suburban                           0.4636      0.069      6.726      0.000       0.329       0.599
author:low_knowledge                      0.0730      0.089      0.824      0.410      -0.101       0.246
author:mid_knowledge                     -0.0260      0.083     -0.315      0.753      -0.188       0.136
author:gender                            -0.1400      0.072     -1.939      0.052      -0.282       0.002
==============================================================================
Omnibus:                        1.255   Durbin-Watson:                   1.478
Prob(Omnibus):                  0.534   Jarque-Bera (JB):                1.032
Skew:                          -0.151   Prob(JB):                        0.597
Kurtosis:                       3.094   Cond. No.                         34.1
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

v. BCT (436919) vs NQS (579220)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     average_problem_accuracy   R-squared:                       0.322
Model:                                  OLS   Adj. R-squared:                  0.255
Method:                       Least Squares   F-statistic:                     129.4
Date:                      Thu, 28 Jul 2022   Prob (F-statistic):           8.81e-65
Time:                              11:13:59   Log-Likelihood:                 18.905
No. Observations:                       135   AIC:                            -11.81
Df Residuals:                           122   BIC:                             25.96
Df Model:                                12
Covariance Type:                        HC1
========================================================================================================
                                           coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                    0.4379      0.060      7.258      0.000       0.320       0.556
author                                   0.0950      0.088      1.079      0.281      -0.078       0.268
pretest_avg_problem_accuracy             0.0882      0.061      1.449      0.147      -0.031       0.208
student_std_attempted                    0.2674      0.431      0.620      0.535      -0.578       1.113
student_std_attempted_before_support    -0.0773      0.306     -0.253      0.800      -0.677       0.522
gender                                   0.0821      0.049      1.686      0.092      -0.013       0.178
low_knowledge                           -0.2602      0.074     -3.523      0.000      -0.405      -0.115
mid_knowledge                           -0.1675      0.057     -2.939      0.003      -0.279      -0.056
rural                                    0.1346      0.035      3.821      0.000       0.066       0.204
suburban                                 0.3033      0.042      7.235      0.000       0.221       0.385
author:rural                             0.1184      0.053      2.221      0.026       0.014       0.223
author:suburban                         -0.0233      0.065     -0.361      0.718      -0.150       0.103
author:low_knowledge                     0.0208      0.096      0.216      0.829      -0.168       0.210
author:mid_knowledge                     0.0711      0.125      0.567      0.571      -0.175       0.317
author:gender                           -0.2832      0.113     -2.516      0.012      -0.504      -0.063
==============================================================================
Omnibus:                        3.557   Durbin-Watson:                   1.446
Prob(Omnibus):                  0.169   Jarque-Bera (JB):                3.073
Skew:                          -0.278   Prob(JB):                        0.215
Kurtosis:                       3.487   Cond. No.                     5.70e+16
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The smallest eigenvalue is 1.33e-31. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Appendix C: Author Code to Author Identifier

Author codes are a sequence of three or four characters to uniquely identify a *star-author* in the ASSISTments platform without the need for an identifier. The identifiers are linked below for easy reference to other ASSISTments papers which use the same *star-authors* and to the data above.

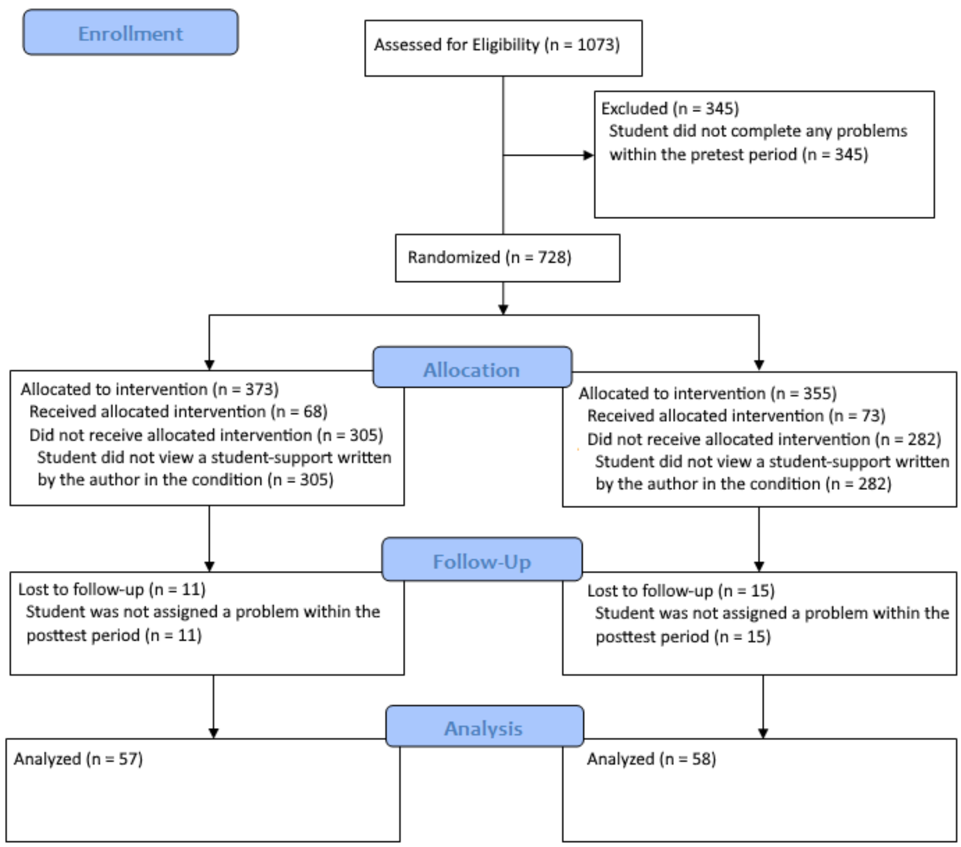
Author Code	Author Identifier
TBD	460571
UETT	255574
NQS	579214
BTS	579216
TAD	460570
EGS	579215
BCT	436919
UEVT	436143
TCD	460572
BGS	579218
LQS	579217
NQS	579220
STD	483200
DMS	578209
UZS	579212
JXS	578208
LLS	294187
OTS	578207
IOT	488160
KTT	485865

Appendix D: Consort Data Flow Plan for Dataset A

CONSORT

TRANSPARENT REPORTING of TRIALS

CONSORT 2010 Flow Diagram



Bibliography

1. Adams, D.M., McLaren, B.M., Durkin, K., Mayer, R.E., Rittle-Johnson, B., Isotani, S., Van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior 36,* 401–411.
2. Feng, M. & Heffernan, N.T. (2006). Informing teachers live about student learning: Reporting in the ASSISTment system. Technology Instruction Cognition and Learning 3(1/2), 63.
3. Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470-497.
4. Lee, J. (2012). Cumulative learning and schematization in problem solving. Universität Freiburg (2012).
5. McLaren, B.M., van Gog, T., Ganoe, C., Karabinos, M., Yaron, D. (2016). The efficiency of worked examples compared to erroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computers in Human Behavior 55*, 87–99.
6. Patikorn, T., & Heffernan, N. T. (2020, August). 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). https://doi.org/10.1145/3386527.3405912.
7. Prihar, E., Botelho, A.F., Jakhmola, R., Heffernan, N.T. (2021). *Assistments 2019-2020 school year dataset.* https://doi.org/10.17605/OSF.IO/Q7ZC5, osf.io/q7zc5
8. Prihar, E. & Gonsalves, M. (2021). *Assistments 2020-2021 school year dataset.* osf.io/7cgav
9. Prihar, E., Patikorn, T., Botelho, A., Sales, A., Heffernan, N. (2021). Toward personalized students' education with crowdsourced tutoring. In: *Proceedings of the Eighth ACM Conference on Learning@Scale*, 37–45. L@S '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3430895.3460130.
10. Razzaq, L.M. & Heffernan, N.T. (2009, June). To tutor or not to tutor: That is the question. In: AIED. pp. 457–464.
11. Whitehill, J. & Seltzer, M. (2017). A crowdsourcing approach to collecting tutorial videos–toward personalized learning-at-scale. In *Proceedings of the Fourth ACM Conference on Learning@Scale*, 157–160.
12. Wood, D., Bruner, J.S., & Ross, G. (1976). The role of tutoring in problem solving. *Child Psychology & Psychiatry & Allied Disciplines.*
13. Prihar, E., Haim, A., Sales, A., & Heffernan, N. (2022, June). Automatic Interpretable Personalized Learning. In *Proceedings of the Ninth ACM Conference on Learning@Scale* (pp. 1-11). Winner of Best Dataset Award. https://doi.org/10.1145/3491140.3528267
14. Team, R. C. (2013). R: A language and environment for statistical computing. https://www.R-project.org/.
15. Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0, https://CRAN.R-project.org/package=pwr.

16. Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... & Busick, M. D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. *National Center for Special Education Research*. https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf
17. Prihar, E., Syed, M., Ostrow, K., Shaw, S., Sales, A. & Heffernan, N. (2022). Exploring Common Trends in Online Educational Experiments. *Proceedings of the 15th International Conference on Educational Data Mining,* 27–38. https://doi.org/10.5281/zenodo.6853041
18. Krichevsky, N., Spinelli, K., Heffernan, N., Ostrow, K., & Emberling, M. R. (2020). E-TRIALS (Doctoral dissertation, Worcester Polytechnic Institute). https://core.ac.uk/download/pdf/343944397.pdf