An M-Estimator for the Survivor Average Treatment Effect

A Major Qualifying Project (MQP) Report
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements
for the Degree of Bachelor of Science in

Mathematical Sciences,
Economics

By:
Lora Dufresne

Project Advisors:
Adam Sales
Alexander Smith

March 6, 2023

This report represents work of WPI undergraduate students submitted to the faculty as evidence of a degree requirement. WPI routinely publishes these reports on its website without editorial or peer review. For more information about the projects program at WPI, see http://www.wpi.edu/Academics/Projects
Abstract

This report implements principal stratification to determine the treatment effect of an educational technology, DragonBox, on middle school algebraic learning. The approach proposes a new solution to the common problem of attrition in the educational sector through the use of principal scores. Findings were modest but indicate the technology may have a positive impact on student learning. This work considers the mathematical trade-off of employing principal stratification and its resultant policy implications.
Acknowledgements

Thank you to Erin Ottmar, Associate Professor of Psychology and Learning Sciences at Worcester Polytechnic Institute, for providing the dataset used in this project. Her generosity has made the work done in this project possible. I would like to extend additional gratitude to Adam Sales, Assistant Professor of Mathematical Sciences at Worcester Polytechnic Institute, and Alexander Smith, Associate Professor of Social Science and Policy Studies, for their guidance as advisors throughout this project.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>ii</td>
</tr>
<tr>
<td>Executive Summary</td>
<td>1</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>6</td>
</tr>
<tr>
<td>2 Background</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Classical Approach to Observing Causal Effects</td>
<td>10</td>
</tr>
<tr>
<td>2.2 Problem</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Relevant Dataset: Educational Technologies on Algebraic Understanding</td>
<td>12</td>
</tr>
<tr>
<td>2.4 Applicable Methods to Solve the Problem</td>
<td>13</td>
</tr>
<tr>
<td>2.4.1 Principal Stratification</td>
<td>14</td>
</tr>
<tr>
<td>2.4.2 M-Estimation</td>
<td>17</td>
</tr>
<tr>
<td>2.4.3 GEEPERs</td>
<td>17</td>
</tr>
<tr>
<td>2.5 Findings from Previous Trials</td>
<td>19</td>
</tr>
<tr>
<td>2.6 Discussion</td>
<td>20</td>
</tr>
<tr>
<td>3 Methodology</td>
<td>22</td>
</tr>
<tr>
<td>3.1 Data Overview</td>
<td>23</td>
</tr>
<tr>
<td>3.1.1 Dataset</td>
<td>23</td>
</tr>
<tr>
<td>3.1.2 Data Application</td>
<td>24</td>
</tr>
<tr>
<td>3.2 Assumptions</td>
<td>25</td>
</tr>
<tr>
<td>3.2.1 Randomization Assumption</td>
<td>25</td>
</tr>
<tr>
<td>3.2.2 Monotonicity Assumption</td>
<td>25</td>
</tr>
<tr>
<td>3.3 Procedure</td>
<td>26</td>
</tr>
<tr>
<td>3.3.1 Addressing Attrition</td>
<td>26</td>
</tr>
<tr>
<td>3.3.2 Bayes Theorem</td>
<td>27</td>
</tr>
<tr>
<td>3.3.3 Standard Errors</td>
<td>30</td>
</tr>
<tr>
<td>4 Results</td>
<td>31</td>
</tr>
<tr>
<td>4.1 Attrition</td>
<td>32</td>
</tr>
<tr>
<td>4.2 Treatment Effect</td>
<td>36</td>
</tr>
<tr>
<td>4.2.1 Standard Errors</td>
<td>44</td>
</tr>
<tr>
<td>4.3 Interpretation of Results</td>
<td>45</td>
</tr>
<tr>
<td>5 Policy Implications</td>
<td>46</td>
</tr>
<tr>
<td>5.1 Pros and Cons of Principal Stratification</td>
<td>47</td>
</tr>
<tr>
<td>5.2 Costs</td>
<td>48</td>
</tr>
<tr>
<td>5.2.1 Cost of Implementation</td>
<td>48</td>
</tr>
</tbody>
</table>
List of Tables

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The four types of principal strata of interest on the post-treatment variable $S_i$, indicating whether student $i$ has completed the posttest or not.</td>
</tr>
<tr>
<td>2</td>
<td>Summary of datasets</td>
</tr>
<tr>
<td>3</td>
<td>Pretest mean score comparison between those students who completed the posttest and those who did not</td>
</tr>
<tr>
<td>4</td>
<td>Pretest mean score comparison by treatment effect between those students who completed the posttest and those who did not</td>
</tr>
<tr>
<td>5</td>
<td>Attrition rates</td>
</tr>
<tr>
<td>6</td>
<td>Attrition rates by treatment</td>
</tr>
<tr>
<td>7</td>
<td>Attrition rates: virtual vs. in-person students</td>
</tr>
<tr>
<td>8</td>
<td>datDragBAU data description</td>
</tr>
<tr>
<td>9</td>
<td>Summary of datDragBAU, a subset of dat03 containing the following covariates</td>
</tr>
<tr>
<td>10</td>
<td>Logit model selection based on AICc</td>
</tr>
<tr>
<td>11</td>
<td>AIC logit model of best fit summary</td>
</tr>
<tr>
<td>12</td>
<td>Summary of the predicted probabilities</td>
</tr>
<tr>
<td>13</td>
<td>OLS model selection based on AICc</td>
</tr>
<tr>
<td>14</td>
<td>Summary of the best fit OLS model with and without the probability term</td>
</tr>
<tr>
<td>15</td>
<td>Difference between treatment effect estimates</td>
</tr>
<tr>
<td>16</td>
<td>Summary statistics of BS standard errors</td>
</tr>
<tr>
<td>17</td>
<td>Standard deviations and 95% confidence intervals for BS standard errors</td>
</tr>
</tbody>
</table>
List of Figures

1. The simple structure of a randomized controlled trial . . . . . . . . . . . . . . . . . . 10
Executive Summary

This report aims to determine the treatment effect for an educational technology, Drag-onBox, while developing a new method for addressing high attrition rates in the educational sector.

Introduction

Many middle school students struggle to learn basic algebraic principles, skills that are necessary for further mathematical advancements. Educational technologies that can aid learning have grown rapidly in popularity within school systems. With the presence of the COVID-19 pandemic accelerating reliance on technology, there are a plethora of options available for educators to choose from. This can make it difficult for educators to determine which technologies will have the greatest impact on student learning outcomes. This project aims to determine the effect of implementing an educational technology in the school system on algebraic learning. A randomized field trial and associated dataset provided by Erin Ottmar, Associate Professor of Psychology and Learning Sciences at Worcester Polytechnic Institute, proved relevant in studying in attempts to solve this issue. This trial was conducted in the midst of the COVID-19 pandemic in which attrition was a serious problem. A focus of this project is to implement a new method for addressing high attrition rates in randomized controlled trials. It is evident that the issue of high attrition rates in educational trials is not isolated to this study, but is present across almost all trials in the educational sector [1]. The hope is that the extreme example of attrition studied in this project may be successful, ultimately proposing a new approach to addressing high attrition rates.

Background

Randomized controlled trials (RCTs) are a common approach to identify causation between variables. This setup makes it possible to estimate average potential outcomes. When analyzing an RCT, it is common to split the sample population into subgroups and estimate separate effects for each group. When subgroup membership is a function of a post-treatment variable where it is affected by the treatment as well, there is said to be post-treatment bias [2]. Attrition bias is a type of post-treatment bias that occurs when there is a loss of participants in the sample population.
before the trial is finished such that they do not complete the trial. Attrition undermines the usual method of RCTs by skewing the causal effects.

Principal stratification is a framework that can solve the issue of attrition. It is possible to adjust for post-treatment variables by thinking of treatment effects for each principal strata, the subgroups of the sample population based on the potential outcomes [3]. M-Estimation can be applied to principal stratification. M-estimators are solutions of the vector equation $\sum_{i=1}^{n} \Psi_i(Y_i, \theta) = 0$ [4]. This approach uses assumptions to yield the average treatment effect, $ATE = E(Y_i(1)) - E(Y_i(0))$, including the Randomization, Stable Unit Treatment Value Assumption (SUTVA), and Monotonicity Assumption. The Randomization Assumption ensures treatment assignment is random. The Monotonicity Assumption assumes that in situations where there is attrition, there is no group that would remain under Control but attrit under Treatment [5]. Another approach, generalized estimating equations for principal effects using regressions (GEEPERs), incorporate principal scores into an M-Estimator for the mixture model. Principal scores as defined by Jo and Stuart 2009 are the probability of being in a principal strata [6].

Under randomization, a model for principal scores can be estimated using data from the treatment group and later extend an application to the control group [2].

Methods

In the dataset provided by Erin Ottmar, Associate Professor of Psychology and Learning Sciences at Worcester Polytechnic Institute, a total of 52 seventh-grade mathematics teachers and their students from 11 middle schools were recruited for the study. 10 of these schools were in-person, while one was a virtual academy. These teachers taught a total of 190 mathematics classes to 4092 students. The investigation involving student attrition rates in this report focuses on the pool of 3271 students from 9 schools. This pool was further split into smaller groups, divided based on student random treatment assignment. The control is Business as Usual (BAU), and the treatments are three educational technologies: DragonBox, From Here to There, and ASSISTments.

For those students randomly assigned to DragonBox and Business as Usual, there were four subgroups of interest: Always-Posttest (AP), Never-Posttest (NP), BAU-Posttest (BP), and Dragon-Posttest (DP). The Randomization Assumption and the Monotonicity Assumption were
invoked for this project. Under monotonicity it is assumed there are no students who would attrit under BAU but complete the trial under DragonBox. Thus, we are left with three subgroups of interest: Always-Posttest, Never-Posttest, and BAU-Posttest.

A logistic regression was run to predict whether a student would complete the posttest based on relevant independent variables. One logistic regression was run for both BAU and DragonBox. Binned error plots and AIC tests were used to aid in the decision of which independent variables to include in the regressions. The likelihood ratio test was used to assess the goodness of fit of models based on their ratio of likelihoods. Bayes Theorem was used to yield results for the principal scores as a ratio of predicted probabilities for students in the Always-Posttest strata (AP) and the subset of students who had an observed posttest (OP). The probabilities are predicted using logistic regression, where the principal score is $Pr(AP|OP) = \frac{Pr(AP)}{Pr(OP)}$. Once the probabilities are estimated, the goal is to find the treatment effect estimate using OLS regression. A treatment effect found without principal stratification is calculated for comparison purposes.

The standard errors were calculated with the bootstrap method. The method mimics the process of random sampling by choosing $N$ subsets of the population of the dataset with replacement, creating an infinite pool of selection. This report takes 1000 samples are taken from the population with replacement.

**Results**

It was found that the attrition rates for all treatments were approaching 50%, with DragonBox having the highest attrition rate at 51.7% when considering all students randomly assigned to the technology.

Including principal scores, the first OLS regression yielded a treatment effect $\hat{\theta}_{AP} = 0.864$. This indicates that in a test out of 10 points, DragonBox may improve a student’s score by 0.864 points relative to BAU. To demonstrate the effectiveness of the predictive probabilities, the model was then run without principal stratification. Without the principal scores, $\hat{\theta}_{OP} = 0.445$, indicating that in a test out of 10 points, DragonBox may improve a student’s score by 0.445 points.

The 95% confidence intervals for the AP and OP methods were (0.105, 1.252) and (0.144, 0.777) respectively. Determining which treatment effect, $\theta_{AP}$ or $\theta_{OP}$, is the better option is not
obvious, as there is a bias-variance trade-off between the two methods. Moving forward with $\theta_{AP}$ means the treatment effect is consistent but has greater variance in the results. $\theta_{OP}$ results in a smaller variance but may be affected by bias. Choosing which treatment effect to use depends on the priorities of those interpreting the results.

There is reason to believe that the Randomization Assumption holds. The Monotonicity Assumption, however, may not be as applicable as once thought. There was appropriate motivation to invoke the assumption, however it may be the case that it is not true in the context of DragonBox and BAU. The risk of the assumption being violated increases the likelihood that the findings are not as accurate as they could be.

**Policy Implications**

This project examines the effects of student technology in the classroom. Implementing DragonBox could replace the addition of an extra teacher at the school, saving the school the cost of the salary that would otherwise have been paid to the teacher. DragonBox costs on average $45,300 per year for a school, while middle school teachers made a median salary of $61,320 in 2021 [7]. This means implementing DragonBox would be about $15,000 less than the cost of an additional teacher. This trade-off would save schools money while implementing technology that improves learning.

School districts pay an average of $310 per Chromebook for their students [8]. For schools that do not have equal access to technology, introducing DragonBox with equal access to computers would require a $235,000 investment every 4 years. With the current funding offered to schools, the average school district would not be able to afford this expense. Along with monetary cost, educators must also consider the time cost for teachers associated with implementing a new technology. One can expect that the implementation time-frame would take time away from educators, as they need to be equipped with the knowledge of the old curriculum as well as the new methods. Once teachers have learned the new methods, however, they would have more available time for planning while DragonBox assists with student learning. High attrition rates can also be seen as a cost to the education system, as attrition rates of close to 50% are likely to cause major questioning among policymakers.

Competitors to DragonBox such as From Here to There have better pricing options and
similar teaching strategies. It is uncertain if DragonBox will stand the test of time against its competitors, however there is promise as the technology is well renowned for its successes in the educational sector. The expansion of AI is into education also highlights the possible fragility of the longevity of DragonBox, as AI has begun to assist student learning [9].

Conclusions

The methods outlined in this report were successful in implementing principal stratification. AIC tests were beneficial in finding the best-fit models for both the logit model for the predictive probabilities and the OLS model for the treatment effect. It was revealed with 95% confidence that DragonBox does indeed have a positive effect on posttest scores. There is, however, room for improvement. The predicted probabilities were fairly noisy and may have affected the accuracy of the results. More sound results could result by looking at which students would have completed the posttest under Business as Usual but not DragonBox. With more time, the standard errors could also be determined using the “sandwich” formulas.

Subsequent work could be done to attempt to find a method outside of logistic regression to find the predictive probabilities. Future work could also focus on gathering better data that reflects the average middle school in the United States, making the results of the study more applicable for policy implementations. Examining student previous knowledge and the treatment effects split by knowledge levels could be a focus with this data: Students could be separated into those with high, medium, and low pretest scores. Results from further work such as this would provide greater insight on the benefits of implementing DragonBox in school systems.
1 Introduction

Educational technologies have undergone a rapid transformation in recent years, with a growing number of schools incorporating digital tools and resources into their curriculum. The presence of the COVID-19 pandemic beginning in 2020 accelerated this reliance on technology, where many students adapted to learning in a fully remote environment. With this widespread adoption of new technologies, there is a plethora of options for teachers and administrators to choose from. From game-based applications to instant student feedback, the possibilities for incorporating technology in the classroom are exponentially expanding. Each application has benefits and limitations. With so many options, it can be difficult for educators to determine which technologies will have the greatest impact on student learning outcomes.

Many middle school students struggle to learn basic algebraic principles [10]. These skills are necessary for further mathematical advancements as concepts are often presented in algebraic notation. To address this gap in knowledge, researchers and developers of educational technologies have designed digital tools to support algebraic learning in the classroom [11]. These technologies are designed to engage students and provide them with a different learning method that could provide higher motivation and timely support. This project aims to determine if implementing an educational technology in the school system could improve algebraic learning. A dataset provided by Erin Ottmar, Associate Professor of Psychology and Learning Sciences at Worcester Polytechnic Institute, proved relevant in studying in attempts to solve this issue. It is important to note this trial was conducted in the midst of the COVID-19 pandemic. Due to the uncertainties faced during this period, attrition was a serious problem. The attrition rates for each treatment were almost or over 50%, far over the accepted attrition rate of under 20% [12]. It is common that those who abandon the trial and attrit have similar motivations for leaving [13]. In the case of this study, there is reason to believe the motivation for leaving may be correlated with the treatment.

A focus of this project is to implement a new method for addressing high attrition rates in randomized controlled trials. When attritors share motivations for leaving the trial, post-treatment bias occurs and regular methods such as OLS regression cannot be implemented. Principal stratification has been identified as a satisfactory approach that can correctly adjust for post-treatment bias [3]. The method can produce treatment effects for each principal strata, the subgroups of the
sample population based on the potential outcomes of treatment assignment. Assumptions often must be made to yield results, as well as to increase clarity and simplicity.

The data detailed in this report observed over 3,600 students and was collected from one fully virtual and nine in-person middle schools in Georgia. All schools were within the same school district. Students received nine 30-minute intervention sessions from September 2020 through March 2021 investigating the impact of educational technologies on seventh-graders’ algebraic knowledge. For the purpose of this project, three different educational technologies were studied: DragonBox (12+), From Here to There, and ASSISTments.

DragonBox (12+) is an educational game that introduces advanced algebraic concepts intended for students ages 12 to 17. One of the key principles that makes DragonBox unique is its design. The application was made such that students should not perceive the game to be focused on math. The research findings on the effectiveness of DragonBox, however, are mixed. From Here to There (FH2T) is also an educational game that uses perceptional learning and embodied cognition to improve mathematics proficiency in students. Previous studies support the efficacy of the educational technology, with results indicating students that completed more problems on FH2T had a higher posttest score than those who were in the control group. ASSISTments is an educational technology that uses immediate-feedback techniques rather than a game-based approach. After completing a problem, students receive immediate feedback on their answer before moving on to the next exercise. Teachers can gather data provided by the application and use the results to guide student instruction based on their needs.

Two assumptions were invoked: the Randomization and Monotonicity Assumptions. Both principal stratification and OLS regression were applied and the results compared. Principal stratification shows that in a posttest out of 10 possible points, DragonBox may increase the test score of a student who would complete the posttest under control or treatment by 0.864. OLS regression reveals DragonBox may increase the test score of a student by 0.455. Determining which method yields better results is not obvious, as there is a bias-variance trade-off. If the assumptions are met under principal stratification then we can say that the treatment effect is consistent but has greater variance in the results. OLS regression results in a smaller variance, but may be affected by bias.

It is evident that the issue of high attrition rates in educational trials is not isolated to this study. This is a problem across almost all trials in the educational sector. Attrition is one
of the major methodological problems in longitudinal studies, where most educational trials are conducted as such [1]. The hope is that the extreme example of attrition studied in this project may be successful, ultimately proposing a new approach to addressing high attrition rates.

With technology becoming increasingly integrated in educational systems, results such as these may prove to influence policy decisions. The policy implications of these results are discussed, taking into account the pros and cons of the methods, costs, and future technological innovations. It was determined although DragonBox may improve algebraic learning, the costs of implementation including capital availability, time, and high attrition rates are not likely to permit incorporating the technology into school systems. With a plethora of educational technologies available, the longevity of DragonBox as a superior option also comes into question. The saturated market provides many substitutes readily available that may be more feasible for schools to implement.

This report first gives a background on randomized controlled trials, attrition, and the methods to address the biases resulting from attrition. The methodology outlines the procedure used for the project, while the results chapter applies these methods to the dataset. Policy implications are discussed along with the significance of the findings in the context of the educational sector.
2 Background

Experiments are often run in an attempt to better one’s understanding of real-world behavior. They are set up to find causal relationships between variables of interest. Experiments employed by researchers are successful in identifying causal relationships rather than just correlation, but are not always unbiased. When bias is identified, these common approaches are untrustworthy and no longer yield accurate conclusions. This report will examine alternative approaches to finding causal relationships when usual methods cannot be applied due to bias caused by attrition.
2.1 Classical Approach to Observing Causal Effects

In order to understand the methods in this report, it is important to first understand the usual process of causal inference. What to do when usual methods cannot be applied will then be explored.

Researchers are interested in causal effects, which is when one variable changes because another has changed. Unfortunately, when one observes the world, one often sees correlation without any evidence that one variable has caused another. Multivariate regression can be used to control for some factors while focusing on the effects of others, but does not prove causation. The only way to ensure one variable caused another is to run an experiment.

Randomized controlled trials (RCTs) are a common approach to confirm causation between variables. As described in [Kendall 2003](#), a RCT is a trial where subjects are randomly assigned to one of two groups: Treatment (T) and Control (C) [14]. The Treatment group is those that receive intervention, while the Control group remains under conventional conditions. The groups are then observed to see if there are differences in their outcomes. Figure 1 displays the simple structure of the RCT.

![Randomized Controlled Trial Diagram](#)

With this setup, one is interested in determining potential outcomes, which are the outcome for an individual under a potential treatment. For an individual in the trial, the causal effect of the treatment is the difference between the potential outcome if the individual receives the treatment and the potential outcome if the individual does not [15].

Potential Outcomes
\( Y_i(1) \) is the potential outcome of person \( i \) if assigned to Treatment

\( Y_i(0) \) is the potential outcome of person \( i \) if assigned to Control

Causal Effect of the treatment

\[
\text{Causal Effect} = Y_i(1) - Y_i(0)
\]

Imagine we have a sample population consisting of \( i = 1, \ldots, N \) individuals. Each individual will either be assigned to Treatment, \( Z_i = 1 \), or Control, \( Z_i = 0 \). The response for the person assigned to either Treatment or Control is \( Y_i \), where \( Y_i(Z) \) is the response for person \( i \) given their treatment assignment \( Z \) \[10\]. We are interested in comparing \( Y_i(1) \) and \( Y_i(0) \), which together embody the entire sample population \[15\].

When performing an RCT, it is common to split the sample population into subgroups as to be able to make comparisons between them. When subgroups are based on pre-treatment covariates the process is straightforward. These subgroups can be observed for treatment effects for a subset of the population \[2\]. For the purpose of our paper, we will consider average treatment effects (ATE).

Treatment Effects

Average Treatment Effect

\[
\text{ATE} = E(Y_i(1)) - E(Y_i(0))
\]

2.2 Problem

When conducting a randomized controlled trial, it is possible to encounter problems that cause bias, which can cause the measured treatment effect to be systematically different from the true treatment effect. This report will focus on post-treatment bias, which transpires when a subgroup membership is a function of a post-treatment variable where it is affected by the treatment as well \[2\]. With this, there may be no causal interpretation to observe.

Attrition bias is a type of post-treatment bias. It occurs when there is a loss of participants in the sample population before the trial is finished such that they do not complete the trial. These individuals’ outcomes of interest cannot be measured due to those participating dropping out. These
outcomes are not “missing”, but rather are defined on a different sample space separate from the observed outcomes [5].

Attrition undermines the usual method of RCTs by skewing the causal effects. The relationship between the dependent and independent variables may be affected and the full effect of the treatment may not be observed. Attrition rates can be different depending on the group: Treatment or Control. It is common that those who abandon the trial and attrit have similar motivations for leaving versus remaining [13]. When the motivation for members of the trial leaving is correlated with the treatment, the post-treatment bias occurs and regular methods cannot be implemented. Other biases exist such as selection bias and performance bias, but focus will remain on attrition bias for the purpose of this report.

2.3 Relevant Dataset: Educational Technologies on Algebraic Understanding

A relevant dataset provided by Erin Ottmar is used in this report. An educational trial, this study focused on examining if educational technologies help middle school students learn algebra. The dataset has one Control, Business as Usual (BAU), and three Treatments: DragonBox (12+), From Here to There, and ASSISTments. The study measures student learning through a pretest and a posttest, both out of 10 possible points.

DragonBox (12+) is an educational game that introduces advanced algebraic concepts intended for students ages 12 to 17 [17]. The application is lauded as one of the best educational games for teaching algebra [18]. Each problem is constructed with the goal of isolating a box that contains a dragon. Isolating this box is equivalent to solving an algebraic equation for \(x\). One of the key principles that makes DragonBox unique is its design. The application was made such that students should not perceive the game to be focused on math. The research findings on the effectiveness of DragonBox, however, are mixed. Some researchers have found the application shows significant gains in learning, while others determined students showed no improvement.

From Here to There (FH2T) is also an educational game that uses perceptual learning and embodied cognition to improve mathematics proficiency in students [19]. The design of the game is based on previous research that suggests learning mathematics is inherently perceptive
and visual representations of notation can affect how students ultimately learn mathematics [20]. FH2T uses color and notation to direct student attention to develop automatic routine for algebraic reasoning. Students are also able dynamically manipulate the equations by dragging and tapping on-screen, allowing students to see their actions’ effects on the expressions. Previous studies support the efficacy of the educational technology, with results indicating students that completed more problems on FH2T had a higher posttest score than those who were in the control group [21]. The findings suggest this application may be successful in increasing student algebraic learning and performance.

ASSISTments is an educational technology that uses immediate-feedback techniques rather than a game-based approach [22]. The problems given to students online resemble those in traditional mathematics textbooks. The problems are presented one at a time to students with the option to request hints if students feel they need supplementary support on a problem. After completing a problem, students receive immediate feedback on their answer before moving on to the next exercise. Teachers can then gather the data provided by the application and use the results to guide student instruction based on their needs [23].

This educational trial was conducted during the COVID-19 pandemic. The students were offered an in-person and online option and were permitted to change methods throughout the trial if pertinent. Because of the uncertainty at this time, attrition was a serious problem. With attrition rates approaching 50%, there is a very high likelihood of attrition bias in this study if usual methods were to be invoked. This report focuses on the educational technology with the highest attrition rates: DragonBox. The following section will outline some applicable methods to solve the critical issue of attrition in the educational sector.

2.4 Applicable Methods to Solve the Problem

Principal stratification is a framework that can solve the issue of attrition. The following section will examine this technique and the particular approaches that can extend from it, namely M-Estimation and GEEPERs.
2.4.1 Principal Stratification

Principal stratification has been identified as a satisfactory approach to mitigate attrition bias. The approach, used in causal inference, adjusts results for post-treatment covariates. The method is focused in the potential outcomes framework and is considered a generalization of the local average treatment effect (LATE). As outlined in Angrist et al. 1996 in the potential outcomes framework “the causal effect of a treatment on a single individual or unit of observation is the comparison (e.g., difference) between the value of the outcome if the unit is treated and the value of the outcome if the unit is not treated” [16]. The target of estimation, called the estimand, is the average causal effect of the treatment, where the average causal effect is defined as the average difference between treated and untreated outcomes across all units in a population.

Recall the sample population established in Section 2.1. An incorrect approach to adjust for post-treatment bias is to make a comparison between the distributions $P\{Y_i(1) \mid S_i = s, Z_i = 1\}$ and $P\{Y_i(0) \mid S_i = s, Z_i = 0\}$ where $S_i$ is a binary measure of the post-treatment variable, taking on a value of 1 or 0. Frangakis and Rubin 2002 notes that comparisons such as these cannot be considered causal effects. Since $S_i$ is a post-treatment variable, it itself can be affected by the treatment [3]. This means that there are potential outcomes for $S_i$: $S_i(1)$ and $S_i(0)$ where $S_i = S_i(Z)$. So both the outcomes $S$ and $Y$ are potential outcomes, as only one version of them can ever be observed. The version that can be observed is that that occurred under assigned treatment $Z$. The unobserved outcome versions are under unassigned treatments $z \neq Z$.

Using principal stratification we can correctly adjust for post-treatment bias. It is possible to adjust for post-treatment variables by thinking of treatment effects for each principal strata, the subgroups of the sample population based on the potential outcomes [3]. Comparisons of outcomes can then be made within a particular strata. For the RCT as described in this report, there are four strata based on treatment assignment and the posttest: Always-Posttest, BAU-Posttest, Dragon-Posttest, and Never-Posttest. $S_i(z)$ is the posttest completion indicator for student $i$. For DragonBox, $z = DB$ and for Business as Usual, $z = BAU$. $S_i(z) = 1$ indicates a student has completed the posttest, while $S_i(z) = 0$ indicates attrition from the trial. Table 1 below outlines what defines each strata in regard to attrition.
Table 1: The four types of principal strata of interest on the post-treatment variable $S_i$, indicating whether student $i$ has completed the posttest or not.

<table>
<thead>
<tr>
<th>$S_i(B) = 1$</th>
<th>$S_i(B) = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Always-Posttest:</strong> completes the posttest under assignment of DragonBox and Business as Usual</td>
<td><strong>BAU-Posttest:</strong> completes the posttest under assignment of Business as Usual and attrits under DragonBox</td>
</tr>
<tr>
<td><strong>Dragon-Posttest:</strong> completes the posttest under assignment of DragonBox and attrits under Business as Usual</td>
<td><strong>Never-Posttest:</strong> attrits and does not complete the posttest under assignment of DragonBox or Business as Usual</td>
</tr>
</tbody>
</table>

Note that the only strata where effects can be defined is Always-Posttest. This is the group that completes the trial under both Treatment and Control.

Assumptions are often made in principal stratification for the purpose of simplicity and increased clarity. Common assumptions include the Randomization Assumption, Stable Unit Treatment Value Assumption (SUTVA), and Monotonicity Assumption.

Throughout the literature reviewed for the purpose of this report, the definition of the Randomization Assumption has widely varied from paper to paper. The purpose of this assumption is to ensure treatment assignment is random. This is not always explicitly noted as an assumption because it is almost always automatically satisfied by the use of a randomized controlled trial. Although not always noted, it is important to use this assumption. Ding et al. [2011] remark that when the Randomization Assumption is implemented, it is certain all treatment assignment is random. This reduces bias that could lead to false conclusions [24].

The Stable Unit Treatment Value Assumption (SUTVA) assumes there is no interference between units. This means the potential outcomes of one individual do not depend on the treatment status of another individual. This also means there is only one version of treatment [24]. Notation as written in Zhang and Rubin [2003] is as follows [5]:
if \( Z_i = Z'_i \) then \( S_i(Z) = S_i(Z') \)

This assumption only holds in certain trials depending on the nature of the experiment. In medical studies with noninfectious diseases, it is likely that there is no interference between individuals in the sample population. One member being assigned to treatment and dying will not affect whether another member assigned to the same treatment will survive or die \[^3\]. If the disease is infectious, there is a high likelihood of interference through the passing of disease from one individual to another. In the case of educational studies, however, it is likely that there is indeed interference. The assumption will be violated by any interaction between students that may influence their decisions. For example, if one student was going to attend class but was persuaded to not attend from a student who was not planning on attending, the SUTVA assumption would be violated.

The Monotonicity Assumption assumes that in situations where there is attrition, there is no group that would remain in the trial under Treatment but attrit under Control \[^5\]. For this project, this means there are no students who would attrit under BAU but complete the trial under DragonBox. Those who attrit under BAU are assumed to be in the Never-Posttest (NP) strata. There are three subgroups of interest remaining: Always-Posttest (AP), Never-Posttest (NP), and BAU-Posttest (BP). Notation as written by Angrist et al. 1996 is as follows \[^16\]:

\[
S_i(1) \geq S_i(0) \text{ for all } i = 1, ..., N
\]

In an educational context, this assumption formalizes the notion that a student under Treatment will not be academically worse off than their peers assigned to Control.

The AP group holds great importance because it is the only strata that is observed under both Treatment and Control. In the case of attrition, the treatment effect is only defined for \( S(0) = S(1) = 1 \), true for the AP strata. The usual method for finding treatment effects using OLS regression works when only two strata exist: Always-Posttest and Never-Posttest. In this circumstance, the observation of the posttest does not differ based on treatment assignment. This is the case when \( S(0) = S(1) \) for all students.
2.4.2 M-Estimation

M-Estimation can be applied to principal stratification to address post-treatment bias, including the issue of attrition. The method commonly performs robust regressions adjusted for outliers. M-Estimation works by attempting to reduce the influence of outliers by replacing the squared residuals in Ordinary Least Squares (OLS) regression by another function of the residuals \[25\]. According to Stefanski & Boos, the M-estimator \( \hat{\theta} \) satisfies

\[
\sum_{i=1}^{n} \Psi_i(Y_i, \hat{\theta}) = 0
\]

where \( Y_i \) represents the \( i \)th datapoint \[4\]. They are consistent for \( \theta_0 \) if \( E[\Psi(Y_i, \theta_0)] = 0 \).

M-Estimation provides a specific approach to mitigate attrition bias. If there are equations that can map to the parameters used, then the method can be applied.

Generalized estimating equations for principal effects using regressions (GEEPERs) incorporate principal scores into an M-Estimator for the mixture model. Principal scores as defined by Jo and Stuart 2009 are defined as

\[
e(X_i) = Pr(S_i(1) = 1 | X_i^S)
\]

where \( X \) is a vector of covariates \[6\]. Under the Randomization Assumption, \( Pr(S_i(1) = 1 | X_i^S, Z_i) = Pr(S_i(1) = 1 | X_i^S) \). With this, a model for principal scores can be estimated using data from the treatment group and later extend an application to the control group \[2\].

2.4.3 GEEPERs

GEEPERs estimates rely on correct specification of regression functions, but do not make assumptions about the distribution of regression errors as in mixture modeling \[2\]. Instead, GEEPERs rely on the Covariate Ignorability Assumption. This assumption states that given a set of covariates \( X_i \), the potential outcomes that would be realized under Treatment \( Y_i(1) \) and Control \( Y_i(0) \) are independent of treatment assignment \( S_i \) \[26\]. That is, the assumption satisfies

\[
E[Y_i(0)|X_i^S, S_i(1)] = E[Y_i(0)|S_i(1)] = \mu_C^0 \text{ or } \mu_C^1
\]
Under this assumption, covariates are not informative of the mean of $Y_i(0)$ within principal strata. Note that it is rare that the Covariate Ignorability Assumption will actually be plausible. However, when given a set of principal scores and under the assumptions of strong Monotonicity and Randomization, the two principal effects for Treatment and Control can be estimated simply by using OLS regression. This means that the estimating equations for the GEEPERs method are equivalent to the estimating equations fit by OLS. That is, for a set of estimating equations

$$ \sum_{i=1}^{n} \tilde{\Psi}_i = 0$$

where

$$ \tilde{\Psi}_i = \begin{bmatrix} (1 - Z_i)[Y_i - \mu_0^C - e(X_i)(\mu_1^C - \mu_0^C)] \\ (1 - Z_i)[e(X_i)Y_i - e(X_i)\mu_0^C - e(X_i)^2(\mu_1^C - \mu_0^C)] \\ Z_i[Y_i - \mu_0^T - S_i(1)(\mu_1^C - \mu_0^C)] \\ Z_i[S_i(1)Y_i - S_i(1)\mu_0^T - S_i(1)^2(\mu_1^C - \mu_0^C)] \end{bmatrix}$$

If the covariate part of model of interest is correct, the Covariate Ignorability Assumption is no longer necessary and the covariates can simply be put into the regression. Lemma 1 from Sales (2020) implies that $E[\Psi_i] = 0$, and thus the estimating equations $\sum_{i=1}^{n} \tilde{\Psi}_i = 0$ are equivalent to the equations fit by OLS \cite{2}. If we define $R_i$ such that

$$ R_i = Z_iS_i + (1 - Z_i)e(X_i) = \begin{cases} e(X_i) & Z_i = 0 \\ S_i(1) & Z_i = 1 \end{cases} $$

(1)

then for $E[Y_i|Z_i, R_i] = \beta_0 + \beta_1R_i + \beta_2Z_i + \beta_3Z_iR_i$ that is fit with OLS with coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$,

$$ \hat{\tau}_0 \equiv \hat{\beta}_2 $$

$$ \hat{\tau}_1 \equiv \hat{\beta}_3 - \hat{\beta}_2 $$

are M-estimators and principal effects can be estimated with a simple regression \cite{2}.
2.5 Findings from Previous Trials

Throughout the literature reviewed, trials were completed with the framework outlined in the previous section. Most experiments used the most common principal stratification approach, while fewer implemented the approach of M-Estimation.

Principal stratification was used in the vast majority of the papers reviewed for this report, including Angrist et al. 1996, Zhang and Rubin 2003, Ding et al. 2011, Stefanski and Boos 2002, and Feller et al. 2017. Angrist et al. 1996 deals with the issue of non-compliance, where people do not act according to their assignment. The paper focuses specifically on the randomized Vietnam draft lottery with the goal of evaluating the effect of serving in the military on health outcomes [16]. Although this report does not focus on non-compliance, the process of mitigating the bias applies to attrition as well. The IV estimand approach was implemented, showing how the estimand can provide straightforward causal interpretation in the potential outcomes framework. The exclusion restriction and Monotonicity Assumption form the core of the IV estimand approach, where the exclusion restriction states that $Y(Z, S) = Y(Z', S)$ for all $Z, Z'$ and for all $S$. It was found that those who were not eligible for the draft had a much lower probability of serving in the military and a slightly lower probability of death and suicide. Hence, serving in the military had a negative causal effect on death and suicide rates.

Zhang and Rubin 2003 and Ding et al. 2011 focused on bias from truncation by death and their methods can be directly applied to attrition bias. The first paper utilized principal stratification in its most basic form. The key example considered two high school educational programs, Treatment (T) and Control (C), and examined the causal effects of the school program on final test scores of those who graduated. Those who dropped out and did not take the final test were “truncated.” The observation is that 30% of students assigned to Treatment graduate with an average test score of 700. However, these students are a mixture of AP and BP as it cannot be observed what happened if they were not assigned Treatment. Applying principal stratification, the causal effect is contained large sample bound [-200, 50]. The latter paper utilizes causal parameters of interest which does not need the assumption of a mixture normal distribution for outcomes. The experiment is interested in the causal effect of two medications, docetaxel and estramustine (DE) and mitoxantrone and prednisone (MP), on the health related quality of life (HRQOL) one year after receiving treatment.
based on survival time. Among the LL group, it is concluded that the DE and MP treatments have similar effects on the HRQOL one year after taking the treatments.

M-Estimation was also implemented in the paper [Stefanski and Boos 2002] [1]. Although Stefanski and Boos 2002 focused on the identification of M-Estimation, an application based on National Basketball Association (NBA) playoffs of 2000 was provided. This example tested the equality of success probabilities of Shaquille O’Neal’s free throws across games. Using the basic M-Estimation approach, it was determined there is some evidence that Shaquille’s foul-shooting probabilities are not constant across games. The effect is not strong, however, and the results are very sensitive to game 14 where all 9 of free throws taken were made.

Sales 2022 implemented the GEEPERs method with a randomized educational field experiment: Middle school mathematics on a computer program [2]. There were four conditions: Treatment in which immediate, informative error messages were provided after entering a wrong answer, Business as Usual which retained the same environment, and two computer programs From Here To There and DragonBox. Although there were high numbers of attrition, it was ignored for the purpose of the experiment. Following the GEEPERs method, results show there were significantly higher posttest scores for students who used From Here to There and DragonBox when compared to the Active Control condition.

### 2.6 Discussion

The majority of the literature reviewed for this report focuses heavily on identification. The authors discuss in great detail whether the methods are possible with the data and outline all equations and assumptions necessary to determine such. However, most do not pursue estimation with the same attention. Once it is known that the method is possible, how does one go about applying it to a finite sample?

Estimation is important because it is how the methods are applied to real-life datasets. Without a focus on estimation, all that has been learned about principal stratification, M-Estimation, and GEEPERs is for nought. We would only establish what we want to know without obtaining estimates. Although Sales 2022 had a focus on estimation, attrition was ignored [2]. There is an open problem where one could use GEEPERs to estimate the Survivor Average Treatment Effect.
This report is looking to solve this lack of exploration by placing estimation at the center of focus. The methods described in Section 2.4 will be applied to an educational dataset with attrition bias. The GEEPER approach will be utilized to find causal effects from the same dataset as the example in [Sales 2022]. The following sections will provide thorough insight into the estimation strategy based on this specific dataset, walking through why certain assumptions are implemented and how we arrive at our results.
3 Methodology

The central objective of this project is to identify the treatment effects of the educational technologies for middle school students in mathematics, specifically those for DragonBox as compared to Business as Usual. To complete this objective, the data set was cleaned and examined. Following initial logistic regressions to identify attrition, estimated probabilities were calculated with the use of Bayes Theorem. A final regression can be written and run with the potential outcomes of DragonBox and Business as Usual, yielding the treatment effect of the DragonBox technology.
3.1 Data Overview

This section first describes the dataset as it was received and follows with a discussion of what subset of data will be used for the purpose of this project.

3.1.1 Dataset

The dataset was provided by Erin Ottmar, Associate Professor of Psychology and Learning Sciences at Worcester Polytechnic Institute. The data was gathered from a large, suburban district in the Southeastern United States in the summer of 2020.

Overall, a total of 52 seventh-grade mathematics teachers and their students from 11 middle schools were recruited for the study. 10 of these schools were in-person, while one was a virtual academy. These teachers taught a total of 190 mathematics classes to 4092 students. Before random assignment, one in-person school dropped out of the study. Students enrolled in resource settings were also not included in the study. This resulted in 10 schools, 37 teachers, 156 classes, and 3612 total students. Random assignment occurred at this stage with the group of schools, teachers, classes, and students described. Following random assignment, one in-person school withdrew from the study as well. This resulted in 9 schools, 34 teachers, 143 classes, and 3271 total students. Out of these 3271 students, 1850 completed both the pretest and posttest assignments for the study. The investigation involving student attrition rates in this report focuses on the pool of 3271 students from 9 schools.

Given that this trial was conducted during the COVID-19 pandemic, the school district offered students and their families a choice of classroom format: either total in-person learning or total online asynchronous learning. Random assignment occurred across both classroom formats in order to account for both learning conditions. Within all classrooms, all teachers were able to use all educational technologies. Based on previous mathematics scores, students were grouped into block sets of five. From these blocks, students were assigned to one of the four options: Business as Usual(20%), DragonBox(20%), From Here to There(40%), and ASSISTments(20%). A greater proportion of FH2T was assigned because the goal of the original experiment run with this dataset was to test the efficacy of FH2T compared to other educational technologies. Since most of the block sets of five grouped students within the same classroom, we will treat random assignment as
blocked within classrooms. In randomized trials, blocking ensures precision in an experiment. As outlined by the General Services Administration, "blocking involves creating homogeneous subsets of the experimental units, then randomly assigning treatments within those subsets, called blocks" \[28\]. It is common practice to block based on quantities that are strongly prognostic of the outcome of interest. This ensures student study conditions were equivalent in regard to their teacher and classroom environment.

### 3.1.2 Data Application

The subsection of data that this report will be focusing on is the pool of 3271 students from 9 schools. The students from the school dropped before random assignment do not have assignments and therefore are not included in the attrition. The school that dropped after random assignment is also not considered when considering attrition in the study because it was not the students' choice on whether to stay or leave. Keeping these students in the study would cause potential bias in the results.

From *The Impacts of Three Educational Technologies on Algebraic Understanding in the Context of COVID-19*, it is known that DragonBox had the largest estimated impact on student success with higher posttest scores \[29\]. However, many individuals had trouble installing the technology. Because for the purpose of this report there is a specific focus on the attrition rates and treatment effect of DragonBox versus the other educational technologies, the pool of 3271 students from 9 schools was further split into smaller groups. The data was divided based on student random treatment assignment. For those students randomly assigned to DragonBox and Business as Usual, there were four subgroups of interest: Always-Posttest (AP), Never Posttest (NP), Business as Usual Posttest (BU), and DragonBox Posttest (DP). It is important to note that these four subgroups are not observed, as we can only witness the outcome of a student according to the treatment in which they were assigned. The subgroups are important to distinguish as they will be used to define the treatment effect of interest. Table 1 in the Background outlines the subgroups. Notice that a treatment effect on the posttest is only well-defined for the Always-Posttest group. This group will therefore be targeted when determining the treatment effect.
3.2 Assumptions

In order to carry out the investigation, two major assumptions must be implemented: The Randomization Assumption and the Monotonicity Assumption. Without these two assumptions, this study would not be possible as we would not be able to determine the subgroups as previously defined. Refer to page 17 of the report for general definitions of these assumptions. The following subsections describe the application of these assumptions in the context of the subset of data that was used to obtain results.

3.2.1 Randomization Assumption

Recall that the Randomization Assumption makes certain that all treatment assignment is random. For this particular dataset, randomization occurred at the student level. As discussed in the dataset section, blocking within classrooms ensured equivalence of student conditions. This method of classroom blocking allows for easy implementation when it comes to estimating fixed effects. It is common to put a dummy variable for classroom to differentiate the blocks of students. The specifics of the random assignment involved a little more intricacy, but the overall method can be thought of as block random assignment to ensure separation, balance in sample sizes, and equal environment.

3.2.2 Monotonicity Assumption

In order to see if the Monotonicity Assumption is valid, the data was split into groups based on treatment assignment. Attrition rates per group were then calculated in R. Attrition rates for those assigned to DragonBox were much higher than the rest of the educational technologies and the control, Business as Usual. The Monotonicity Assumption for this dataset assumes that those who attrit under BAU will attrit under DragonBox. It follows that:

\[ S_i(D) \geq S_i(B) \text{ for all } i \text{ students} \]

Monotonicity is more plausible given the observed attrition rates than if attrition were higher under Business as Usual. Because of the much higher attrition rates for those assigned to DragonBox, it may be appropriate to use the assumption, which rules out the possibility of having
Dragon-Posttest students in the trial. Under monotonicity, there are no students who would attrit under BAU but complete the trial under DragonBox. Those who attrit under BAU are assumed to be in the Never Posttest (NP) strata. Thus, we are left with three subgroups of interest: Always-Posttest (AP), Never-Posttest (NP), and BAU-Posttest (BP).

Invoking the Monotonicity Assumption allows subgroup assignment for a student given their observed attrition and treatment assignment. We can say that those who take the posttest under DragonBox assignment are AP, those who attrit under DragonBox are either NP or BP, those who take the posttest under DragonBox assignment are AP, those who attrit under Business as Usual are either AP or BP, and those who attrit under Business as Usual are NP.

3.3 Procedure

With the end goal of finding a treatment effect for the educational technology DragonBox, the subject of attrition must first be addressed. Then Bayes Theorem can be invoked to estimate probabilities.

3.3.1 Addressing Attrition

As mentioned in the motivation for invoking the Monotonicity Assumption, the first task was to examine the attrition rates for each treatment assignment, verifying that DragonBox has the highest attrition rate of all.

Once decisions were final on the assumptions, the next task involved dealing with missing data in pretest, posttest, and numerical covariates. This process involved removing values recognized as “NA” in R from all of the variables with missing data. Two logical variables were created, one for those who have completed the pretest and one for who have completed the posttest. This latter variable, hasPosttest, is the dependent variable for the logistic regression that will be run to predict whether a student will complete the posttest. It reads TRUE if the student completed the posttest and FALSE otherwise. Dummy variables were created for student race and student school. There were n-1 of these dummies for each non-numerical variable of interest, where n = # of choices of the categorical data. For data that was numerical but missing values, the NA was replaced by the mean of the observed data. In this case, we are not interested in the covariates themselves. The purpose
in removing the missing values is to use them to predict attrition, so imputation is a viable method.

After all missing values were dealt with, the data was then split into groups based on treatment assignment. With these groups, it was then possible to run a logistic regression to predict whether a student would complete the posttest based on relevant independent variables. One logistic regression was run for both BAU and DragonBox.

Binned error plots were used to aid in the decision of which independent variables to include in the regressions. Numerical variables could be graphed versus the dependent variable HasPosttest and the datapoints were placed into “bins” such that one can observe if there is a resultant pattern or not. Variables that were not numerical were chosen based on motivation from previous literature.

The likelihood ratio test assesses the goodness of fit of two models based on their ratio of likelihoods. One of these models is a subset of the other, lacking independent variables that the other has. It worked as follows:

\[ H_0 : \text{The full model and the subset model fit the data equally well} \]
\[ H_1 : \text{The full model fits the data significantly better than the subset model} \]

In total, 7 models were run. The first included independent variables for having the pretest, if the student is male, if the student is considered gifted, dummy variables for student race, number of unexcused days off from school, the pretest score for students that took the test, the time the student took to complete the pretest, and dummy variables for the school that the student attends. The log was taken for the time the student took to complete the pretest to create a better fit. The model was constructed as follows:

\[ Pr(\text{hasPosttest}_i) = \logit^{-1}(\beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \alpha_{s[i]}) \]

where \( \text{hasPosttest} \) is a logical variable, \( X_j \) are the covariates of interest, and \( \alpha_{s[i]} \) is a binary indicator of the school of student \( i \). The likelihood ratio was run on smaller and smaller subsets of this model until the null hypothesis is rejected, indicating that the previous model was the best fit.

### 3.3.2 Bayes Theorem

After completing the logistic regressions, Bayes Theorem was needed to yield results on the treatment effects. Bayes Theorem is the probability of an event based on prior knowledge of
conditions that might be related to the event. Mathematically for events A and B, the theorem can
be written as

\[ Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)} . \]

Note that \( Pr(B) \) cannot equal 0 as it is in the denominator. For our research question,
we are interested in predicting posttest scores. The observed posttest scores, \( Y_i \), in BAU are either
part of the Always-Posttest or BAU-Posttest strata. We write the mean as the weighted average of
the probabilities, with

\[ E(Y|BAU) = Pr(AP)\mu_{BAU}^{AP} + Pr(BP)\mu_{BAU}^{BP} \]

where \( \mu_{BAU}^{AP} \) is the control mean for AP and \( \mu_{BAU}^{BP} \) is the control mean for BP. From these control
means, we are interested in decomposing the mean of the observed outcomes. This is why we have
to condition on \( OP \); otherwise including another term, \( Pr(NP)\mu_{NP} \), would be necessary.

The first method used the notion that we can find \( Pr(AP) \) with the logistic regression.
\( Pr(AP|X) \) was estimated with the logistic regression described previously. These are the students
that fall into Always-Posttest (AP) in the DragonBox treatment assignment. What we do not have
is \( Pr(AP|X, OP) \) where OP is the event that the posttest is observed and X is the covariates. Let
us assume all probabilities are conditional on X such that we need \( Pr(AP|OP) \). Recall the equation
for Bayes Theorem: For some A and B, \( Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)} \). Thus we can rewrite \( Pr(AP|OP) \)
as \( \frac{Pr(OP|AP)Pr(AP)}{Pr(OP)} \). Since \( Pr(OP|AP) = 1 \), we are left with \( Pr(AP|OP) = \frac{Pr(AP)}{Pr(OP)} \). \( Pr(OP) \)
is the probability of the posttest being observed, specifically in the BAU condition. This is what
was estimated in the BAU logistic regression model. The DB logistic regression model estimated
\( Pr(AP) \), where \( Pr(OP) = 1 - Pr(NP) \).

The second method fits one logistic regression to both the AP and BP groups and uses a
dummy variable for conditions interacted with the covariates. It is then possible to either include
or exclude covariates from the model based on the coefficients of their interaction terms, not their
main effects. To find the probabilities, a fake column of treatment assignments is created within
the dataset, assigning all students to either DragonBox or Business as Usual. The model is fit to
the original(real) data, with the fake data used as a shortcut to getting the estimated probabilities
\( Pr(hasPostest|DB) = Pr(AP) \) and \( Pr(hasPostest|BAU) = 1 - Pr(NP) = Pr(OP) \). With this, we
are able to find the two estimated probabilities as was done in the first method, with \( Pr(AP|OP) = \frac{Pr(AP)}{Pr(OP)} \).

Once the probabilities are estimated, the goal is to find the treatment effect estimate. The potential outcomes for DragonBox and Business as Usual are as follows: For the BP subgroup, let \( B = 1 \). For the AP subgroup, let \( B = 0 \). Let \( \pi = Pr(AP|OP) = \frac{Pr(AP)}{Pr(OP)} \). We have the following potential outcomes for posttests under BAU and DragonBox

\[
E[Y(B)|X, \pi, B] = \alpha_0 + \alpha_1 B + \beta' X
\]

\[
E[Y(B)|X, \pi] = \alpha_0 + \alpha_1 (1 - \pi) + \beta' X
\]

and

\[
E[Y(D)|X, B] = \beta_0 + \beta' X
\]

These two equations can be combined into one large equation that can be estimated with an OLS regression. For the DragonBox condition, \( \pi = Pr(AP|OP, X) = 1 \) because every student with an observed posttest in DragonBox is AP by the Monotonicity Assumption. For the Business as Usual condition, \( \pi = \frac{Pr(AP)}{Pr(OP)} \) is estimated from logistic regressions. The combined equation predicts student posttest scores as follows

\[
Y = \alpha_0 + \alpha_1 (1 - \pi) + \theta Z + \beta' X + \epsilon
\]

where \( Z = 0 \) for BAU and \( Z = 1 \) for DragonBox

Once the regression is run, the treatment effect can be found by finding

\[
E[Y|Z = 1, B = 0] - E[Y|Z = 0, B = 0]
\]

We yield

\[
\alpha_0 + \alpha_1 (1 - \pi) + \theta + \beta' X - (\alpha_0 + \alpha_1 (1 - \pi) + \beta' X)
\]

\[= \theta\]

where \( \theta \) is the treatment effect estimate.
3.3.3 Standard Errors

The standard errors were calculated with one of the most widely used tools in statistics to estimate uncertainty – the bootstrap. As defined by Jonathan Bartlett, the bootstrap method "is a computational resampling technique for finding standard errors (and in fact other things such as confidence intervals), with the only input being the procedure for calculating the estimate (or estimator) of interest on a sample of data" [30]. The method mimics the process of random sampling by choosing $N$ subsets of the population of the dataset with replacement, creating an infinite pool of selection. The sample of data that is taken is treated as the entire population. This satisfies the definition of a standard error, where the standard error reflects the variability between estimates obtained by repeatedly taking samples from the population.

For the bootstrap method, 1000 samples are taken from the population with replacement. For each of these 1000 samples, the data is treated as the entire population. First, the probabilities are estimated. Then the point estimates are calculated and an OLS regression is run to find a treatment effect. The standard deviation of all of the treatment effects found across the 1000 samples serves as a sufficient standard error measure. We know that the treatment effect from our original sample population is $\hat{\theta}$. Let the estimated treatment effects from the bootstrap samples be $\hat{\theta}_n$ for $n = \{1, 2, ..., 1000\}$. Then the bootstrap standard error for $\hat{\theta}$ is

$$SE(\hat{\theta}) = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (\hat{\theta}_n - \bar{\theta})^2}$$

where $\bar{\theta}$ is the mean of the estimates across the $N$ bootstrap samples. For $N = 1000$ we have that

$$SE(\hat{\theta}) = \sqrt{\frac{1}{999} \sum_{n=1}^{1000} (\hat{\theta}_b - \bar{\theta})^2}$$

where $\bar{\theta} = \frac{1}{1000} \sum_{n=1}^{1000} \hat{\theta}_n$

The bootstrap method was run in R and was successful in yielding an estimate of a standard error. In the next section, the results of applying this methodology are presented.

A link to the full project code can be found in the footnote\footnote{https://github.com/ldufres/m-estimator} below.
4 Results

In this section, the methods previously described are implemented. The results are grouped into three main subsections: Attrition, Treatment Effect, and Interpretation of Results. The importance of probability predictions is exemplified with a comparison of results with and without these predictions. The procedure in the Methodology serves as a road map for yielding the following results.
4.1 Attrition

Aligning with the data description outlined in section 3.1.1, the full dataset was filtered. The filtering of the dataset was done in steps following the timeline of school and student attrition. Table 2 shows the sorting process including the number of schools, teachers, classes, and students in each dataset.

<table>
<thead>
<tr>
<th>Data Description</th>
<th>Dataset</th>
<th>Schools</th>
<th>Teachers</th>
<th>Classes</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>full dataset</td>
<td>dat00</td>
<td>11</td>
<td>52</td>
<td>190</td>
<td>4092</td>
</tr>
<tr>
<td>school 1 drops</td>
<td>dat01</td>
<td>10</td>
<td>48</td>
<td>173</td>
<td>3715</td>
</tr>
<tr>
<td>resource students drop</td>
<td>dat02</td>
<td>10</td>
<td>37</td>
<td>156</td>
<td>3612</td>
</tr>
<tr>
<td>school drops post-treatment assignment</td>
<td>dat03</td>
<td>9</td>
<td>34</td>
<td>143</td>
<td>3271</td>
</tr>
<tr>
<td>students who took pre and post</td>
<td>dat04</td>
<td>9</td>
<td>34</td>
<td>127</td>
<td>1850</td>
</tr>
</tbody>
</table>

Table 2: Summary of datasets

The dataset dat00 is the full number of schools, teachers, classes, and students recruited for the study. Dat01 reflects the first school dropping out of the study. Dat02 filters dat01, removing those students who are enrolled in resource. Dat03 reflects the previous filtering and in addition drops the school that declined participation following the pretest assessment. The final dataset, dat04, includes only the subset of students from dat03 who completed both the pretest and posttest. Dat03 and dat04 will be used for comparative purposes throughout this section. There exists another subgroup of dat03, those students who took the pretest but did not take the posttest. These are students who attrited.

Both the pretest and posttest scores for this study are out of 10 possible points. The tests are 10 questions, with each question worth one point. Table 3 shows the pretest scores of two groups of students: Those who completed both the pretest and posttest and those who only completed the pretest. These mean scores include students from all four treatment assignments – Business as Usual, DragonBox, From Here to There, and ASSISTments.

Observe that the pretest mean score for students that completed the posttest is higher than the pretest mean score those students who attrited and did not complete the posttest. This result was expected, as students who performed well on the pretest may have more incentive to also take the posttest. These students are also most likely better performing students in general.
Students Mean Pretest Score
completed posttest 4.842
did not complete posttest 4.405

Table 3: Pretest mean score comparison between those students who completed the posttest and those who did not

These subgroups, those who completed both the pretest and posttest and those who only completed the pretest, were then further subdivided based on treatment assignment. Dividing the students by treatment assignment allows one to see which educational technology yielded the greatest pretest mean score. The results are in Table 4 below.

<table>
<thead>
<tr>
<th>Students</th>
<th>BAU Mean</th>
<th>DB Mean</th>
<th>FH2T Mean</th>
<th>ASSIST Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>completed posttest</td>
<td>4.702</td>
<td>4.963</td>
<td>4.805</td>
<td>4.937</td>
</tr>
<tr>
<td>did not complete posttest</td>
<td>4.310</td>
<td>4.640</td>
<td>4.297</td>
<td>4.451</td>
</tr>
</tbody>
</table>

Table 4: Pretest mean score comparison by treatment effect between those students who completed the posttest and those who did not

There is a wide variation in the pretest scores by treatment. When observing students who completed the posttest, BAU yielded the lowest mean score. Dragonbox had the highest, followed closely by ASSISTments and FH2T. DragonBox and BAU are separated by 0.261, indicating students assigned to DragonBox yielded a 2.61% higher score. When observing students who did not complete the posttest, FH2T yielded the lowest mean score. DragonBox again had the highest score followed by ASSISTments and BAU. DragonBox and BAU are separated by a slightly wider margin of 0.330, indicating students assigned to DragonBox resulted in a posttest that had a 3.3% higher score. The variation between scores was greater for those students who did not complete the posttest than for those who did.

DragonBox yielded the highest mean pretest score for both those students that completed the posttest and those who did not. For the group that did not complete the pretest, the DB mean dropped by a smaller margin than all other educational technologies. This observed DragonBox advantage served as motivation to pursue a treatment effect for DragonBox as opposed to the other technologies.
It is crucial to address attrition before attempting to find a treatment effect. Dat03 and dat04, as shown in Table 2, are used to find and compare attrition rates for students that completed both the pretest and posttest versus those who only completed the posttest. The equation for calculating attrition rates is

\[
\text{attrition rate} = \frac{\# \text{ of students left}}{\# \text{ of students randomized}}
\]

where \# of students left denotes the number of students that finish the trial and \# of students randomized refers to the number of students originally randomly assigned to a treatment. Using this equation, attrition rates were calculated under 2 conditions: those who took the posttest and those who took both the pretest and the posttest. These conditions correspond to subsets of dat03. A subset of dat03 of students who completed the posttest and dat04, a subset of dat03 of students who completed the pretest and the posttest, were treated as the \# of students left, while the \# of students randomized used the total number of students from dat02, 3612, as the final school that was filtered out dropped from the study after randomization occurred.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Attrition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>students who completed pretest</td>
<td>0.461</td>
</tr>
<tr>
<td>all students</td>
<td>0.488</td>
</tr>
</tbody>
</table>

Table 5: Attrition rates

Note from Table 5 that the attrition rate is higher when requiring both a pretest and a posttest score as opposed to only a posttest score. This is logical, as the group of students who took both the pretest and the posttest is a subset of the group of students that took the posttest. Attrition rates were then calculated by treatment, with wide variation resulting. Notably, DragonBox yielded the highest attrition rates for both the group of students who completed the posttest and the group of students who completed both the pretest and posttest. For the first group, the attrition rate for DragonBox was 2.2% higher than the next highest rate under BAU and 4.3% higher than the lowest rate under ASSISTments. For the latter group, the attrition rate for DragonBox was 2.3% higher than the next highest rate under BAU and again 4.3% higher than the lowest rate under
### Attrition Rates

<table>
<thead>
<tr>
<th>Conditions</th>
<th>BAU</th>
<th>DragonBox</th>
<th>FH2T</th>
<th>ASSISTments</th>
</tr>
</thead>
<tbody>
<tr>
<td>students who completed pretest</td>
<td>0.467</td>
<td>0.489</td>
<td>0.451</td>
<td>0.446</td>
</tr>
<tr>
<td>all students</td>
<td>0.494</td>
<td>0.517</td>
<td>0.477</td>
<td>0.474</td>
</tr>
</tbody>
</table>

Table 6: Attrition rates by treatment

Attrition rates of greater than 20% are considered unacceptable for RCTs as bias may be introduced. Attrition rates greater than 30% are indicators that there may be flaws in the trial and proceeding as usual is not advised, as the results may be skewed and affect later decisions such as policy advising. With Table 6 displaying attrition rates approaching 50%, it is obvious that additional methods must be applied in order to ensure that the results are sound. The very high attrition rates in this particular dataset can be attributed to the COVID-19 pandemic. The effects of going virtual can be seen in the table below.

<table>
<thead>
<tr>
<th>Students</th>
<th>Attrition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>virtual</td>
<td>0.557</td>
</tr>
<tr>
<td>stayed virtual</td>
<td>0.597</td>
</tr>
<tr>
<td>in-person</td>
<td>0.390</td>
</tr>
<tr>
<td>stayed in-person</td>
<td>0.382</td>
</tr>
</tbody>
</table>

Table 7: Attrition rates: virtual vs. in-person students

Notice in Table 7 the attrition rate for virtual students is almost 20% higher than for in-person students. The focus on DragonBox and the importance of addressing attrition is emphasized with the attrition rates under DragonBox being the highest by a notable margin. This serves as motivation to move forward with the predictive probabilities using Bayes Theorem as discussed in Section 3.3.2.
4.2 Treatment Effect

To begin finding a treatment effect, a new dataset was created. This dataset, datDragBAU, is a subset of dat03 containing only those covariates relevant to a student’s performance on the pretest and posttest assessments. Binned scatter plots were used in R to find patterns between the logical covariates Pretest and hasPosttest and other numerical covariates that showed promise of relevance. The resulting dataset contains 15 covariates, 12 of which are logical and 3 being numerical. A table outlining variable descriptions is provided in Table 8. A summary of the dataset is shown in Table 9.
<table>
<thead>
<tr>
<th>Statistic</th>
<th>Data Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>posttestScore</td>
<td>sum of the scores on the 10 math knowledge problems in posttest for each student (out of 10 points)</td>
</tr>
<tr>
<td>hasPosttest</td>
<td>logical variable that indicates if a student has taken the posttest</td>
</tr>
<tr>
<td>drag</td>
<td>logical variable that indicates if a student was assigned to DragonBox</td>
</tr>
<tr>
<td>mathscores</td>
<td>sum of the scores on the 10 math knowledge problems in pretest for each student adjusted for missing pretests (out of 10 points)</td>
</tr>
<tr>
<td>pretest</td>
<td>logical variable that indicates if a student has taken the pretest</td>
</tr>
<tr>
<td>times</td>
<td>log of the average time taken to complete 10 problems (in seconds)</td>
</tr>
<tr>
<td>gifted</td>
<td>logical variable that indicates if a student is gifted</td>
</tr>
<tr>
<td>S1</td>
<td>logical variable that indicates if a student attends school 1</td>
</tr>
<tr>
<td>S2</td>
<td>logical variable that indicates if a student attends school 2</td>
</tr>
<tr>
<td>S3</td>
<td>logical variable that indicates if a student attends school 3</td>
</tr>
<tr>
<td>S4</td>
<td>logical variable that indicates if a student attends school 4</td>
</tr>
<tr>
<td>S5</td>
<td>logical variable that indicates if a student attends school 5</td>
</tr>
<tr>
<td>S6</td>
<td>logical variable that indicates if a student attends school 6</td>
</tr>
<tr>
<td>S8</td>
<td>logical variable that indicates if a student attends school 8</td>
</tr>
<tr>
<td>S9</td>
<td>logical variable that indicates if a student attends school 9</td>
</tr>
</tbody>
</table>

Table 8: datDragBAU data description

An additional covariate, worry, indicated a student’s level of worry before the pretest was
Table 9: Summary of datDragBAU, a subset of dat03 containing the following covariates taken. This covariate is included in the AIC models due to promising relevance but proved to not improve the fit of any model. Because of its lack of functionality, worry was not included in the final datDragBAU dataset.

In order to determine an optimal logistic regression, a selection is made based on the Akaike Information Criterion (AIC) method. The AIC method works by evaluating how well a model fits the data it was generated from \[31\]. It can be used to compare different models to ultimately determine which is the best fit for the data. The AIC establishes the best model as the one that explains the greatest amount of variation using the fewest possible independent variables, where the AIC is calculated using the number of independent variables in the model and the maximum likelihood estimator. The method also penalizes models that use more parameters, avoiding redundancy. This means that if there are two models that explain the same amount of variation, the best-fit model will be that with the fewest parameters. The lower the AIC score, the better fit the model is to the data.

The combinations of covariates in the models included in the AIC model were decided based on logical connections and experimental design. Table 10 shows the results of the AIC model. Table 11 outlines the results of the best-fit model.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>posttestScore</td>
<td>755</td>
<td>4.502</td>
<td>2.838</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>hasPosttest</td>
<td>1,308</td>
<td>0.577</td>
<td>0.494</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>drag</td>
<td>1,308</td>
<td>0.500</td>
<td>0.500</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>mathscores</td>
<td>1,308</td>
<td>4.705</td>
<td>2.457</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>pretest</td>
<td>1,308</td>
<td>0.872</td>
<td>0.334</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>times</td>
<td>1,308</td>
<td>4.207</td>
<td>0.725</td>
<td>1.192</td>
<td>7.603</td>
</tr>
<tr>
<td>gifted</td>
<td>1,308</td>
<td>0.157</td>
<td>0.364</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S1</td>
<td>1,308</td>
<td>0.085</td>
<td>0.279</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S2</td>
<td>1,308</td>
<td>0.042</td>
<td>0.201</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S3</td>
<td>1,308</td>
<td>0.066</td>
<td>0.248</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S4</td>
<td>1,308</td>
<td>0.101</td>
<td>0.301</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S5</td>
<td>1,308</td>
<td>0.099</td>
<td>0.298</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S6</td>
<td>1,308</td>
<td>0.408</td>
<td>0.492</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S8</td>
<td>1,308</td>
<td>0.068</td>
<td>0.252</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>S9</td>
<td>1,308</td>
<td>0.069</td>
<td>0.253</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
interaction terms with drag | K | AICc  | Delta AICc | AICcWt | Cum.Wt | LL   
--- | --- | ------ | ---------- | ------ | ----- | ----- 
times    | 14  | 1423.24 | 0.00      | 0.18   | 0.18  | -697.46 
none      | 12  | 1423.51 | 0.28      | 0.16   | 0.34  | -699.64 
times and gifted | 16  | 1423.54 | 0.30      | 0.16   | 0.50  | -695.56 
times and pretest | 15  | 1424.16 | 0.92      | 0.11   | 0.61  | -696.89 
gifted    | 14  | 1424.39 | 1.16      | 0.10   | 0.71  | -698.03 
pretest   | 13  | 1424.43 | 1.19      | 0.10   | 0.81  | -699.08 
mathscores | 13  | 1424.74 | 1.50      | 0.09   | 0.90  | -699.23 
mathscores.pretest | 14  | 1425.68 | 2.44      | 0.05   | 0.95  | -698.68 
times and worry | 16  | 1427.30 | 4.06      | 0.02   | 0.98  | -697.44 
worry     | 14  | 1427.52 | 4.28      | 0.02   | 1.00  | -699.60 

Table 10: Logit model selection based on AICc

\[ \text{logit}(\text{hasPosttest}) = \beta_0 + \beta_1 \cdot \text{drag} + \beta_2 \cdot \text{Pretest} + \beta_3 \cdot \text{mathscores} + \beta_4 \cdot S1 + \beta_5 \cdot S2 + \beta_6 \cdot S3 + \beta_7 \cdot S4 + \beta_8 \cdot S5 + \beta_9 \cdot S6 + \beta_{10} \cdot S8 + \beta_{11} \cdot S9 + \beta_{12} \cdot \text{times} + \beta_{13} \cdot (\text{drag} \cdot \text{times}) \]

<table>
<thead>
<tr>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasPosttest</td>
</tr>
<tr>
<td><strong>drag</strong></td>
</tr>
<tr>
<td><strong>pretest</strong></td>
</tr>
<tr>
<td><strong>mathscores</strong></td>
</tr>
<tr>
<td><strong>S1</strong></td>
</tr>
<tr>
<td><strong>S2</strong></td>
</tr>
<tr>
<td><strong>S3</strong></td>
</tr>
<tr>
<td><strong>S4</strong></td>
</tr>
<tr>
<td><strong>S5</strong></td>
</tr>
<tr>
<td><strong>S6</strong></td>
</tr>
<tr>
<td><strong>S8</strong></td>
</tr>
<tr>
<td><strong>S9</strong></td>
</tr>
<tr>
<td><strong>times</strong></td>
</tr>
<tr>
<td><strong>drag:times</strong></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
</tr>
</tbody>
</table>

Observations 1,308
Log Likelihood -697.457
Akaike Inf. Crit. 1,422.913

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: AIC logit model of best fit summary

This model contains an interaction term between the covariates drag and times. With
the lowest AICc score, the log(time) that it took those students assigned to DragonBox shows a compelling association with a treatment assignment to DragonBox.

This logistic regression model was then applied to predict the two probabilities of interest, $Pr(\text{AP})$ and $Pr(\text{OP})$. DatDragBAU was split into two subsets, dbDat and bauDat. The dataset dbDat contained all of the students from datDragBAU assigned to DragonBox, while the bauDat contained all of the students from datDragBAU assigned to the control, BAU. The predict() function in R was used to find these probabilities. The function works by predicting values based on the input data. This is why it was crucial that the model was one of best-fit. The R code for $Pr(\text{AP})$ and $Pr(\text{OP})$ is shown below.

```r
prAP <- predict(times, dbDat, type = 'response')

prOP <- predict(times, bauDat, type = 'response')
```

Bayes theorem was then applied

$$\text{prAPO} \leftarrow \frac{\text{prAP}}{\text{prOP}}$$

and the probabilities were successfully predicted. The results are shown in Table 12.

<table>
<thead>
<tr>
<th>Probability</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>prAP</td>
<td>0.02559</td>
<td>0.42634</td>
<td>0.58264</td>
<td>0.56283</td>
<td>0.79666</td>
<td>0.95802</td>
</tr>
<tr>
<td>prOP</td>
<td>0.02997</td>
<td>0.45259</td>
<td>0.61376</td>
<td>0.59088</td>
<td>0.81970</td>
<td>0.96452</td>
</tr>
<tr>
<td>prAPO</td>
<td>0.6916</td>
<td>0.9035</td>
<td>0.9440</td>
<td>0.9368</td>
<td>0.9763</td>
<td>1.1604</td>
</tr>
<tr>
<td>prob</td>
<td>0.6916</td>
<td>0.9035</td>
<td>0.9440</td>
<td>0.9356</td>
<td>0.9763</td>
<td>1.0000</td>
</tr>
<tr>
<td>estprobs</td>
<td>0.00000</td>
<td>0.02369</td>
<td>0.05598</td>
<td>0.06445</td>
<td>0.09647</td>
<td>0.30836</td>
</tr>
</tbody>
</table>

Table 12: Summary of the predicted probabilities

$Pr(\text{AP}|\text{OP}) > 1$ is returned in 4% of cases, a number deemed acceptable after thorough discussion. Because of the possibility of statistical error, it was possible to set these probabilities greater than one exactly equal to one. This set of probabilities was simply called prob. It was then feasible to find the probability that prob did not occur, that is, $1 - Pr(\text{AP}|\text{OP})$. Called estprobs, this step was necessary due to the method applied later to ultimately determine the treatment effect.
\[ estprobs = (1 - \text{prob}) \]

It is now the case that all students assigned to DragonBox should have a probability of 0 for \( estprobs \). To ensure this is the case, the final regression uses \( \text{probsadj0} \), which has the results from \( estprobs \) for those students assigned to BAU and 0 for those assigned to DragonBox.

The AIC method was then used again to determine the best-fit OLS model. The dependent variable is now \( posttestScore \), the posttest score for those who took the posttest. This corresponds to the final goal of a treatment effect of the DragonBox technology on the posttest score results. The results of the method and the model of best fit are both displayed below.

Consider the following as the baseline model:

\[
\begin{align*}
\text{posttestScore} &= \\
\beta_0 + \beta_1 \text{probsadj0} + \beta_2 Z + \beta_3 \text{pretest} + \beta_4 \text{times} + \beta_5 \text{gifted} + \beta_6 S1 + \beta_7 S2 + \beta_8 S3 + \beta_9 S4 + \beta_{10} S5 + \\
&\quad \beta_{11} S6 + \beta_{12} S8 + \beta_{13} S9 + \epsilon
\end{align*}
\]

The names of the models in Table 13 are labeled based on the baseline model.

<table>
<thead>
<tr>
<th>model</th>
<th>K</th>
<th>AICc</th>
<th>Delta AICc</th>
<th>AICcWt</th>
<th>Cum.Wt</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>mathscores</td>
<td>16</td>
<td>3328.97</td>
<td>0.00</td>
<td>1</td>
<td>1</td>
<td>-1648.12</td>
</tr>
<tr>
<td>mathscores, no gifted or times</td>
<td>14</td>
<td>3354.55</td>
<td>25.58</td>
<td>0</td>
<td>1</td>
<td>-1662.99</td>
</tr>
<tr>
<td>mathscores, no gifted</td>
<td>15</td>
<td>3356.58</td>
<td>27.61</td>
<td>0</td>
<td>1</td>
<td>-1662.97</td>
</tr>
<tr>
<td>worryscores, no times</td>
<td>15</td>
<td>3458.75</td>
<td>129.77</td>
<td>0</td>
<td>1</td>
<td>-1714.05</td>
</tr>
<tr>
<td>worryscores</td>
<td>16</td>
<td>3459.97</td>
<td>130.99</td>
<td>0</td>
<td>1</td>
<td>-1713.62</td>
</tr>
<tr>
<td>no times</td>
<td>14</td>
<td>3469.74</td>
<td>140.76</td>
<td>0</td>
<td>1</td>
<td>-1720.58</td>
</tr>
<tr>
<td>baseline</td>
<td>15</td>
<td>3471.16</td>
<td>142.19</td>
<td>0</td>
<td>1</td>
<td>-1720.26</td>
</tr>
<tr>
<td>worryscores, no gifted or times</td>
<td>14</td>
<td>3532.08</td>
<td>203.11</td>
<td>0</td>
<td>1</td>
<td>-1751.76</td>
</tr>
<tr>
<td>worryscores, no gifted</td>
<td>15</td>
<td>3533.88</td>
<td>204.91</td>
<td>0</td>
<td>1</td>
<td>-1751.62</td>
</tr>
<tr>
<td>no gifted or times</td>
<td>13</td>
<td>3552.56</td>
<td>223.58</td>
<td>0</td>
<td>1</td>
<td>-1763.03</td>
</tr>
<tr>
<td>no gifted</td>
<td>14</td>
<td>3554.52</td>
<td>225.55</td>
<td>0</td>
<td>1</td>
<td>-1762.98</td>
</tr>
</tbody>
</table>

Table 13: OLS model selection based on AICc

\[
\begin{align*}
\text{posttestScore} &= \\
-1.426 + 8.643 \text{probsadj0} + 0.864 \text{drag} + 1.447 \text{Pretest} + 0.507 \text{mathscores} + 0.065 \text{times} + \\
1.337 \text{gifted} + 1.166 S1 + 1.010 S2 + 1.084 S3 + 0.508 S4 + 0.906 S5 + 1.575 S6 + 0.506 S8 + 0.558 S9 + \epsilon
\end{align*}
\]
As opposed to the logistic regression model fit earlier, \textit{modelTrial} does not have interaction terms, which were avoided in this model because they have the potential to greatly increase the standard errors. The results of the model are shown below in Table \ref{table:ols}. The covariate of interest, \textit{drag}, and its corresponding results are in boldface. The standard errors of each covariate indicates the accuracy of the predicted coefficient with respect to the population parameter. The standard error is calculated by taking the square root of the variance of the coefficient.

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>post.total_math_score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(modelTrial)</td>
</tr>
<tr>
<td>probsadj0</td>
<td>8.643** (3.949)</td>
</tr>
<tr>
<td>\textbf{drag}</td>
<td>\textbf{0.864}** (0.246)</td>
</tr>
<tr>
<td>Pretest</td>
<td>1.447*** (0.378)</td>
</tr>
<tr>
<td>mathscores</td>
<td>0.507*** (0.041)</td>
</tr>
<tr>
<td>times</td>
<td>0.065 (0.123)</td>
</tr>
<tr>
<td>GIFTED</td>
<td>1.337*** (0.245)</td>
</tr>
<tr>
<td>S1</td>
<td>1.166 (0.726)</td>
</tr>
<tr>
<td>S2</td>
<td>1.010* (0.531)</td>
</tr>
<tr>
<td>S3</td>
<td>1.084** (0.451)</td>
</tr>
<tr>
<td>S4</td>
<td>0.508 (0.423)</td>
</tr>
<tr>
<td>S5</td>
<td>0.906** (0.431)</td>
</tr>
<tr>
<td>S6</td>
<td>1.575*** (0.386)</td>
</tr>
<tr>
<td>S8</td>
<td>0.505 (0.477)</td>
</tr>
<tr>
<td>S9</td>
<td>0.558 (0.445)</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.426 (0.911)</td>
</tr>
<tr>
<td>Observations</td>
<td>755</td>
</tr>
<tr>
<td>R²</td>
<td>0.427</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.416</td>
</tr>
<tr>
<td>Residual Std. Error</td>
<td>2.168 (df = 740)</td>
</tr>
<tr>
<td>F Statistic</td>
<td>39.387*** (df = 14; 740)</td>
</tr>
</tbody>
</table>

\textit{Note:} *\(p<0.1\); **\(p<0.05\); ***\(p<0.01\)

Table 14: Summary of the best fit OLS model with and without the probability term

Returning to the methods outlined in Section 4.3.2, the covariate \textit{drag} is equal to \(Z\), where \(\text{drag}_i = 1\) if student \(i\) is assigned to DragonBox and \(\text{drag}_i = 0\) if student \(i\) is assigned to Business as Usual. The resulting coefficient of \textit{drag} is equal to the treatment effect estimate \(\theta\). Therefore,

\[
\hat{\theta}_{AP} = 0.864
\]

is the treatment effect estimate. This means that the causal effect of the treatment assignment...
DragonBox is 0.864 points on the posttest score, where the posttest is an exam out of 10 total points.

To demonstrate the effectiveness of the predictive probabilities, the model *modelTrial* was then run without the covariate *probsadj0*. This model was named *adjmodelTrial*. The results of the removal of *probsadj0* is also displayed in Table 14 with the covariate of interest, *drag*, in boldface.

Without *probsadj0*,

$$\hat{\theta}_{OP} = 0.445$$

is the treatment effect estimate. This means that the causal effect of the treatment assignment DragonBox is 0.445 points on the posttest score, where the posttest is an exam out of 10 total points. This is 0.419 less than the treatment effect estimate estimated with *estprobs*. The impact of the predictive probabilities was significant in improving the estimate and reducing potential bias in the model. Table 15 shows the difference in treatment effect estimates in these two methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Treatment Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.864</td>
</tr>
<tr>
<td>OP</td>
<td>0.455</td>
</tr>
</tbody>
</table>

Table 15: Difference between treatment effect estimates

The treatment effect estimated with principal stratification almost doubles that estimated with regular OLS regression.
4.2.1 Standard Errors

To calculate the standard errors, a for loop was used in R. The entire method for finding the treatment effect estimate was run a total of 1000 times, where the group of students for each loop was chosen randomly with replacement. A summary of the findings is shown in Table 16. The beginning of the loop can be found in the Appendix.

<table>
<thead>
<tr>
<th>BS Standard Errors</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd.Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>-0.3106</td>
<td>0.4821</td>
<td>0.6791</td>
<td>0.6788</td>
<td>0.8689</td>
<td>1.7377</td>
</tr>
<tr>
<td>OP</td>
<td>-0.0613</td>
<td>0.3469</td>
<td>0.4536</td>
<td>0.4543</td>
<td>0.5634</td>
<td>0.9374</td>
</tr>
</tbody>
</table>

Table 16: Summary statistics of BS standard errors

The standard deviations and 95% confidence intervals for AP and OP are shown in Table 17 below.

<table>
<thead>
<tr>
<th>BS Standard Errors</th>
<th>Standard Deviation</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.287</td>
<td>(0.105, 1.252)</td>
</tr>
<tr>
<td>OP</td>
<td>0.158</td>
<td>(0.144, 0.777)</td>
</tr>
</tbody>
</table>

Table 17: Standard deviations and 95% confidence intervals for BS standard errors

Notice that AP, the method that used principal stratification, has much higher standard errors. Although not ideal, this does not come as a surprise. In order to address attrition, the Monotonicity Assumption was invoked and students were split into groups based on treatment assignments. The assumption in itself causes some noisiness in the predictions. Paired with the predictive probabilities that follow using Bayes Theorem, attribute noise likely increased when these values were then used in the best-fit OLS model.

Because of the large standard errors, confidence in the significance of the treatment effects is decreased. The mean for the bootstrap treatment effects is very close to the calculated treatment effect for OP, off by about 0.001. In the case of AP, however, the mean for the bootstrap is about 0.185 less than the value calculated previously. This discrepancy means the results from AP must be interpreted with caution. The results are more difficult to interpret, as the 95% confidence interval provides a wide range of values for the treatment effect estimated with principal stratification. It can
be said with 95% confidence that both methods resulted in a positive treatment effect, increasing the scores at a minimum of 0.105 and 0.144 points respectively.

4.3 Interpretation of Results

Determining which treatment effect, $\theta_{AP}$ or $\theta_{OP}$, is the better option is not obvious, as there is a bias-variance trade-off between the two methods. Moving forward with $\theta_{AP}$ means the treatment effect is consistent but has greater variance in the results. $\theta_{OP}$ results in a smaller variance, but may be affected by bias. Choosing which treatment effect to use depends on the priorities of those interpreting the results.

There were two assumptions made for the purpose of this project: the Randomization and Monotonicity Assumptions. After viewing the findings, one can determine whether these assumptions held throughout the process. The Randomization Assumption presumes that all treatment assignment is random. There is no rationale to claim that this assumption does not hold, as the data was gathered through a randomized controlled trial (RCT). The Monotonicity Assumption, however, may not be as applicable as once thought. Although it is the case that more students were to attrit under DragonBox than BAU, the variation between attrition rates is much more considerable for students who stayed virtual compared to students who completed the study in-person. There was appropriate motivation to invoke the assumption, however it may be the case that it is not true in the context of DragonBox and BAU. The risk of the assumption being violated increases the likelihood that the findings are not as accurate as they could be.
5 Policy Implications

In this section, the results are examined to determine what these findings may mean for policy in the educational sector. Along with the more standard considerations of costs, access to technology, ease of installation, and technological advances, data-specific factors such as attrition and standard errors are explored as well. The findings from the previous chapter indicate that principal stratification, while better at navigating the issue of attrition, resulted in much larger standard errors. This raises the question of which method would be more sensible for a school system to use in their decision-making regarding the implementation of DragonBox.
5.1 Pros and Cons of Principal Stratification

The treatment effects, $\hat{\theta}_{AP} = 0.864$ and $\hat{\theta}_{OP} = 0.455$, indicate that a student in the indicated strata will increase their score by $\theta$ when assigned to the treatment DragonBox when compared to the BAU condition in an exam out of 10 possible points.

First, we will consider the use of the principal stratification method. For this dataset, attrition rates vary by treatment effect for these educational technologies. Principal stratification is the optimal choice to navigate RCTs with attrition, especially with the high attrition rate for DragonBox. Under this method, it is possible to identify underlying subgroups and compute causal effects only within the always-posttest (AP) subgroup. We gain the benefit of being able to compare treatments adjusting for post-treatment variables that may yield principle effects, that is, effects within each subgroup. Since the principal strata are not affected by treatment assignment, they can be treated as pre-treatment covariates and it is always the case that the effects for these principal strata are causal effects. Utilizing the results from the principal stratification method will therefore guarantee that the findings are causal. An argument for the implementation of DragonBox can be made based on the knowledge that the posttest scores differed because of the treatment assignment. The strength of this claim, however, is debatable. Because the standard error is so much larger, confidence in the outcome is low. The confidence interval, $(0.105, 1.252)$, indicates that the assignment of DragonBox could increase a student’s posttest score by about 0.105 to 1.252 points with 95% confidence in an exam out of 10. Even so, due to the large standard error, it may be the case that the sample does not closely represent the population. It is difficult to determine if the AP treatment effect estimate of 0.864 is accurate. Even if the estimate is deemed acceptable, principal stratification only addresses those students who never attrit. We are unsure if the causal effect would be the same for those who do not fall into this strata.

There is then the option of forgoing principal stratification, resulting in the observed posttest (OP) treatment effect. This method is less effective at addressing attrition. Not addressing attrition can lead to bias, as the attrition rate under DragonBox was higher than the other treatment assignments. It can also weaken the credibility of the RCT, decreasing confidence in the results of the trial. It is important to note that this method does not guarantee causal effects since not all of the outcomes are observed. This method, however, addresses all students, not just those that never
attrit. Benefits of this method include a significantly smaller standard error, indicating that the sample mean more closely reflects the population mean. The OP treatment effect estimate of 0.455 can therefore be argued as a better estimate when attempting to ensure a small confidence interval.

Ultimately, the decision between the methods and using either the AP or OP treatment effects comes down to preference and what is more important to the individual evaluating the pros and cons of the educational technology. If one would like a guaranteed causal effect that accounts for attrition, then they would likely choose principal stratification. Their goals would need to be more methods based and less focused on the closeness of fit for the population mean, as the trade-off involves greater standard errors. If the fit of the sample to the population is more paramount to a policymaker, they would likely choose to forego principal stratification. It would be easier to make a case for entire school systems to implement or abandon the educational technology if an argument can be made for the sample mean is a close reflection of the population mean. However, a sacrifice is made in the reliance on the effect being causal. There is no longer a guarantee that the results are indeed causal. This could undermine an otherwise promising argument either in favor of the technology.

5.2 Costs

Standard in policy decisions, cost evaluation allows for a more well-informed verdict. Determining the monetary cost of implementing the educational technology and the additional cost of ensuring student access is likely to drive a policy decision in a certain direction. Costs are also not limited to monetary constraints. In this section, opportunity costs are discussed for the consideration of both the schools and the students involved.

5.2.1 Cost of Implementation

Costs of implementation are a major consideration when determining whether to move forward with new policy. In the case of educational technologies such as the one focused on in this report, the cost of a new method can be the final decision-making element, even if the technology itself is proven to be impactful. Intervention that prompts change is often expensive and school systems operate on limited funds. Schools receive the majority of their funding from state and local
sources with little help from the federal government [32]. For technology, local funding is often the main source of capital. Different districts have different funding capacities necessary to cover these costs, often coming up short. As a result, the school systems need to rely on multiple sources in attempts to piece together enough money, and even then often fall short of their needs.

Federal and state funding exists through programs that endorse technology-enabled learning. These programs include the Every Students Succeeds Act (ESSA), Individuals with Disabilities Education Act (IDEA), and E-rate program. Limits include a 15% cap on ESSA funds to support technology and 20% to 90% discounts on internet access only with E-rate[32]. There are limited options in how this money can be spent. Often when schools secure this funding, they are forced to pick and choose what the money will go toward, leading to a difficult decision. Philanthropic gifts and fundraising also exist as an option to increase funds but can be unreliable because of their one-time nature.

DragonBox offers a free 7-day trial. After the 7 days expire, it costs $9.99 per month or $59.99 per year per student for access. For the consideration of large-scale implementation, $59.99 per year per student makes the most financial sense. The dataset from the trial comes from middle schools in the Southeast United States, specifically Georgia. According to Public School Review, the average number of students attending a public middle school in Georgia is approximately 755 students[33]. This means it would cost approximately $45,300 per year for the average middle school in Georgia to implement this new educational technology, a cost that is infeasible and unsustainable for most school districts. However, one must consider that schools often get a discount for high-volume purchases. Although discontinued, DragonBox used to provide a 50% educational discount to schools [34]. It is likely that a similar deal exists for schools to purchase the educational technology in bulk, drastically reducing the yearly cost of DragonBox to around $22,650. This annual cost is much more feasible for the average middle school in Georgia.

For insight on school funding, consider the following from the study Closing America’s Education Funding Gaps by The Century Foundation (TCF). Results indicate that “the United States is underfunding its K-12 public schools by nearly $150 billion annually, robbing more than 30 million school children of the resources they need to succeed in the classroom” [35]. Funding gaps are present in the majority of school systems, meaning the schools are operating at below-average outcomes. The average deficit in the United States is more than $5000 per student. It was found
that almost two-thirds of all public school students face funding gaps. Districts that face funding
gaps are disproportionately made up of low-income, Black, and Latinx students. There is a barrier
to entry for those schools that lack sufficient funding considering implementing the technology will
require an upfront investment for all students. The results of this report are arguably not strong
enough to encourage the yearly cost reported above. Even the most compelling results can be set
aside due to a lack of funds. Educators must weigh the benefits of a slight increase in testing scores
with the cost that is not feasible to continue long-term. It appears that at the moment, DragonBox
is too costly for the average school district.

One may argue that the trade-off of technology versus teachers may be more efficient.
Technology in the classroom has been shown to allow for larger student-to-teacher ratios. Drag-
onBox in this case would be a substitute for a teacher in the classroom. The technology grades
tests automatically and provides instruction to students while they are using it, taking these time
commitments away from teachers. This means that implementing DragonBox could replace the addi-
tion of an extra teacher at the school, saving the school the cost of the salary that would otherwise
have been paid to the teacher. Middle school teachers made a median salary of $61,320 in 2021,
about $15,000 less than the cost of DragonBox per year for the average middle school student body
[7]. Disregarding student access, this trade-off would save schools money while implementing tech-
nology that improves learning. If the technology were to be used at home instead, the technology
is then a compliment to classroom learning. In this case, an additional teacher would need to be
hired and DragonBox would need to be purchased, a situation that costs the school more money
instead of providing savings. Hence, the advantages of the trade-off are exclusive to in-classroom
learning. DragonBox, while technologically significant, cannot be considered a complete substitute
for a teacher. The application lacks the ability to help students who are struggling with the material,
as there is no explanatory feature. Teachers are also able to connect with students on a personal
level to offer reassurance when necessary.

5.2.2 Cost of Student Access

The annual cost of providing DragonBox to the student is not the only setback in implement-
tation. The dataset used for this project examined students who already had access to technology.
When thinking about widespread implementation, it is important to consider the case when total
student access may not be the case. One must consider the cost of ensuring all students have access to a computer or tablet, as the technology cannot be utilized without one of these devices. For students that do not currently have access to these types of devices, the opportunity cost is lower for implementing DragonBox than continuing with the control curriculum. Implementing the educational technology would only be beneficial if all students had equal access. For access in the classroom, pairing implementation with one-to-one distribution could be a viable solution. One-to-one distribution proves one device for every student in the school, regardless of their previous access to devices [36].

According to The Economic Value of Chromebooks for Educational Institutions, school districts pay an average of $310 per Chromebook for their students [8]. For the average Georgia middle school, an additional lump sum of about $235,000 would be necessary to ensure equal access for all students. This would increase the price per student of implementing DragonBox to $369.99 per student, a vast increase from the previous $59.99 per student. These laptops, while durable, are used almost daily by students and require regular updates and replacement. It is common for Chromebooks to be replaced in schools after 4 years of use. The schools that choose to employ the use of DragonBox would therefore be committed to a $235,000 investment every 4 years to keep equal access to the educational technology. With the funding as described above, the average school district simply cannot afford this expense.

Introducing new technology takes time. Educators must consider the time cost for teachers associated with implementing a new technology. According to A Complete Guide to Implementing Tech in Schools, it is uncommon for new curriculum to be implemented all at once [37]. Most school districts tend to phase in the new methods over several years to ensure there is support and expertise necessary for student success. The average time-frame for full implementation of new curriculum in the U.S. school system is two to four years. Considering the DragonBox technology addresses quite similar topics that would already be taught in schools, one can expect that the implementation time-frame would shorten significantly. Even if the period was to half, however, one to two years of unveiling a new technology can take quite a bit of time away from educators, as they need to be equipped with the knowledge of the old curriculum as well as the new methods.

Higher attrition rates can also be seen as a cost to the education system. Attrition rates of close to 50% are likely to cause major questioning among policymakers. Although the new technology
shows it can increase test scores for those who complete the study, what about those students who dropped out of the study and did not complete the posttest? Lack of knowledge in what the effect could be for almost half of students is quite discouraging when contemplating the effectiveness of DragonBox. Further considering that DragonBox had the highest attrition rate among all treatments for both students that completed the pretest and students who did not, this technology may not be worth the investment.

5.3 Technology

One of the major complaints regarding DragonBox is its lack of simplicity in installation. With an online rating of 1.8 stars, users of DragonBox 12+ are writing complaints about the poor browser experience. The program developed by Kahoot DragonBox AS is not widely downloaded in the United States, only having about 5000 downloads. It is possible that with increased time and interest, the browser experience will become less challenging. Kahoot DragonBox AS will likely want to invest its time and money into a program that the education system is increasingly demanding. This requires the implementation paired with the interest that at the moment is severely lacking.

A close competitor to DragonBox, From Here to There (FH2T), is currently free to register and open for educators to use. It works similarly as a gamed-based tool, guiding students through puzzles in an interactive manner. There is evidence that students who solved more problems with FH2T had higher posttest scores and the effect is stronger in those students lacking previous knowledge. The existence of such a similar product emphasizes the flaws of DragonBox and provides users with an easy switch if they so desire. If DragonBox is to succeed long-term, the technology undoubtedly needs to become more aligned with its competitors in terms of pricing, benefit, and ease of use.

As a whole technology is expanding exponentially. The computing and processing capacity of computers is about doubling every 1.5 years and it is predicted that by 2030, 500 billion devices will be connected to the internet. The rapid growth is paralleled in the educational sector. According to the World Economic Forum, the educational technology market specifically is projected to expand to 342 billion dollars by 2025. As of 2023, there are approximately 567,000 educational applications available for download. There is no sign that these educational technologies are slowing in development or release. With so many options readily available for educators to use, the
relevance of DragonBox long-term comes into question. The COVID-19 pandemic heavily boosted the popularity of learning with technology, a method that is remaining applicable to this day. Thus, there is a high likelihood that there will be further technological advances in the educational sector in the near future. Although the systems used in DragonBox is at the forefront of technological learning at the moment, the projected continued exponential growth of the market could quickly deem the educational technology outdated. Paired with the reported browser difficulties that are already plaguing users, it can be argued that the program has multiple substitutes readily available. These proposed substitutes may even be more feasible for schools to implement, as many are free to download.

Looking past current educational technologies, artificial intelligence (AI) remains one of the most sought-after technological advancements [39]. The use of AI is beginning to expand into education, extending past grading student assessments into tools that can assist student learning [9]. With the introduction of AI in education, previously established educational technologies could be deemed obsolete. Machine-based AI is already fairly widespread in education and has the potential to drive down the cost of assessment, something that the current technologies cannot compare to [9]. With the prospect helping students learn, better, faster, and more cost-effectively, the introduction of AI is not only inevitable but arguably a necessary next step.
6 Conclusion

This project aimed to find a treatment effect for the educational technology DragonBox while addressing the problem of attrition. While previous methods tend to ignore attritors or treat them as missing, the methodology outlined in this report accounts for high attrition rates by adjusting for post-treatment bias. Through the use of principal strata, Bayes Theorem, and two major assumptions, a set of principal scores were estimated. With these principal scores, it was possible to estimate the treatment effects with simple OLS regression.

6.1 Key Takeaways

Principal stratification is a useful method that can be applied to randomized controlled trials when faced with addressing high attrition rates. In this project it was possible to apply the Monotonicity and Randomization Assumptions, giving insight into how treatment assignment affected the subgroups of students we called principal strata. The introduction of Bayes Theorem allowed the prediction of probabilities used to find the principal scores. The ratio of these probabilities explained the likelihood that a student posttest would be in the AP strata given that the test was observed. Using a best-fit model with OLS regression, treatment effects were estimated with and without the probability covariate, comparing the effectiveness of principal stratification and highlighting the bias-variance trade-off between principal stratification and the usual OLS method.

The methods applied in this report were able to achieve the goal of finding a treatment effect while accounting for attrition. The treatment effect under principal stratification, $\theta_{AP}$, was a resultant average effect that was causal and consistent. This outcome, however, had a large variance that resulted in large standard errors. The usual method, with the result $\theta_{OP}$, had a much smaller variance but was not guaranteed to be causal or unbiased. This bias-variance trade-off indicates there is not a simple choice regarding which method and resultant treatment effect is the better option. Ultimately the goals of policymakers will determine which method may be employed.

The implications of these results rely not only on the effectiveness of DragonBox but also on the costs schools must bear and the longevity of the technology in the educational sector. Regardless of the significance of the results, schools will not be able to make policy changes if they do not have
the available funds to make implementation a reality. The technology also must be able to hold its own in the educational sector among competition, as there are countless interactive technology options available for students. If DragonBox were to be implemented, policymakers would need to be confident that the technology could be used in curriculum for a minimum span of time.

6.2 Success of the Method

The methods outlined in this report were successful in implementing principal stratification. It was possible to find principal scores through a ratio of estimated probabilities that could then be used in OLS regression to find a treatment effect.

AIC tests were beneficial in finding the best-fit models for both the logit model for the predictive probabilities and the OLS model for the treatment effect. Finding these best-fit models ensured that the results would be as accurate as possible. Although the method using principal stratification resulted in a wider variance, the result is confirmed to be both causal and unbiased. In an RCT, an effect being causal is of the utmost importance. Without the confirmation that a treatment effect is causal, one cannot say with certainty that the effect happened because of the treatment. Bias outcomes can lead to false conclusions, a concern one does not need to worry about with principal stratification.

For both the AP and OP treatment effects, it was revealed with 95% confidence that DragonBox does indeed have a positive effect on posttest scores. The AP method had a minimum effect of 0.105 and a maximum effect of 1.252 and the OP method had a minimum effect of 0.144 and a maximum effect of 0.777 in a posttest out of 10 possible points. The motivations compelling the exploration into how DragonBox could help students learn were valid, as the application does show improvement in student learning measured through exam scores.

6.3 Drawbacks to the Method

In this project, the principal scores were calculated as the ratio of two probabilities: \( Pr(AP) \), the probability that students are in the Always-Posttest strata, and \( Pr(OP) \), the probability that the posttest is observed. With Bayes Theorem applied, we had that
principal scores = \( Pr(\text{AP}|\text{OP}) = \frac{Pr(\text{AP})}{Pr(\text{OP})} \)

Each of these probabilities, \( Pr(\text{AP}) \) and \( Pr(\text{OP}) \), were predicted separately. The predictions of these probabilities were fairly noisy and may have affected the accuracy of the results. When the ratio of these probabilities was calculated, some outcomes included values over 1. This cannot be the case, as a probability has a minimum of 0 and a maximum of 1. The best prediction process resulted in about 4% of these ratios with results over 1. Due to the timeline of the project, these values were simply set equal to 1 and ultimately used in the best-fit OLS model. With more time, it would have been better practice to find another method of calculating the principal scores that obtained values more aligned with theory.

This project also looked at which students completed the posttest under DragonBox and Business as Usual when looking for the treatment effect. While working through the methods, it became apparent that really we want to look at which students would have completed the posttest under Business as Usual but not DragonBox. In the final section of the methodology, the roadmap for how to split students by strata is given. Additional time provided, the focus could have been shifted to this strata specifically. With this focus shift could come results varied from those treatment effects estimated with those students who completed the posttest under Dragonbox and BAU.

The bootstrap method was used to estimate standard errors in this project. Its simplicity provided a straightforward way to estimate the standard errors and confidence intervals from estimators that were more complex. However, there is always going to be noise resulting from the resampling process. With more time, the standard errors could instead be determined using the “sandwich” formulas. Sandwich standard errors traditionally estimate the mean and variance, but can be expanded to estimating the principal scores and variance separately. Finding standard error matrices at each step of the methodology will propagate the uncertainty throughout the procedure, ensuring further precision when accounting for standard error.

### 6.4 Future Work

As outlined in the previous section, a better probability model could have a significant impact on the project. Subsequent work on the topic could be done to attempt to find a method outside of linear regression to find these probabilities. With regard to the methods that were used,
this report focused on two for treatment effects: AP and OP. Although it was not possible within the time frame of this project, future work could investigate how sensitive each method is. These assumptions are not necessarily mutually exclusive but tend to contradict each other. One could explore which is stronger and which is weaker, determining when these assumptions hold. Looking deeper into the assumption comparison between these methods could highlight which results may be best for policy decisions.

For the purpose of this project, the dataset provided was already complete. A major drawback of this dataset was the lack of diversity in the student sample. The trial included mostly white students who attended middle school in the state of Georgia, a sample that does not align with the population of middle school students across the United States. Future work could focus on gathering better data that reflects the average middle school in the United States so as to make the results of the study more applicable for policy implementations. To determine the best sample, one could use the diversity index. The diversity index gives a score ranging from 0 to 1, with a score closer to 1 indicating a more diverse student population. According to the Public School Review, public schools in the United States have an average diversity score of 0.69 [42]. Finding a sample of students that have a diversity score of 0.69 or close to it would create a dataset that better reflects the average middle school.

A topic that was discussed but not explored in this project was examining student previous knowledge. A study focusing on the educational technology From Here to There (FH2T) included effects based on student prior knowledge using pretest scores [19]. The results indicated that students who solved more problems with FH2T had higher posttest scores, and the effect was stronger in students with low prior knowledge. A similar focus could be applied to DragonBox. Students could be separated into 3 subgroups based on their prior knowledge: those with high, medium, and low pretest scores. The treatment effect of the educational technology can then be determined for each subgroup. The results would provide insight into which students DragonBox would benefit the most.
References


FH2T Research Team. From here to there (fh2t). URL https://sites.google.com/view/from-here-to-there/overview/fh2t.


Lauren Decker-Woodrow, Craig Mason, Ji-Eun Lee, Jenny Yun-Chen Chan, Adam Sales, Allison Liu, and Shihfen Tu. The impacts of three educational technologies on algebraic understanding in the context of covid-19. *AERA Open*, Forthcoming.


Appendices

A Beginning of Bootstrap Error Loop

for(i in 1:1000){

bsRows <- sample(1:nrow(datDragBAU),nrow(datDragBAU),replace=TRUE)

bsDat <- datDragBAU[bsRows,]

db_bsDat <- mutate(bsDat,drag=1)

bau_bsDat <- mutate(bsDat,drag=0)

...
}