Principal Score Modeling 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

Author:

Anzhe Tao

Advisor:

Adam Sales

December 15, 2023

Abstract

This project addresses the challenges for estimating effects of treatment in an educational experiment with high attrition rates. The project discusses the study of Principal Stratification, focusing on estimating the Average Treatment Effect of Always Post-test stratum - students who would take the post-test no matter which experimental group they belonged to. Since the previous methods for identifying that stratum resulted in principal scores (probabilities) larger than 1, new methods including Multinomial Regression and Expectation-Maximization algorithm are introduced in this project. In the experiment, middle school students used interventions including to enhance their math skills. The result shows that the treatment had a positive impact on post-test math scores.

Key words: Randomized Controlled Trial (RCT), Attrition, Average Treatment Effect (ATE), Principal Stratification, Multinomial Regression, Expectation-Maximization (E-M) Algorithm

Table of Contents

List of Tables

Executive Summary

The purpose of this project is to determine the treatment effect in an educational experiment, as well as introducing new methods to address the high attrition rates.

Introduction

This project presents a study addressing challenges in middle school mathematics education, focusing on the effectiveness of game-based interventions like From Here to There (FH2T) and DragonBox. The study encountered a significant hurdle with a high attrition rate, prompting the development of an innovative solution. The project introduces a novel approach that combines multinomial regression with the Expectation-Maximization (E-M) algorithm to address attrition challenges.

Traditional methods, such as excluding dropouts or assuming zero post-test scores, are known for introducing biases. The proposed approach enhances the analysis by accurately capturing nuances in student performance and engagement while effectively managing missing data. This integration of E-M algorithm and multinomial regression proves advantageous, offering a statistically sound treatment of attrition issues and providing valuable insights into the effectiveness of game-based education in middle school mathematics. The project's findings have broader implications for improving educational technology in mathematics education, emphasizing the importance of addressing attrition challenges in research methodologies.

Background

Randomized Controlled Trials (RCTs) stand as the gold standard in research design, ensuring unbiased estimates by randomly assigning participants. By randomly assigning participants to treatment and control groups, RCTs ensure unbiased estimates, allowing for clear and reliable comparisons between groups and facilitating the calculation of Average Treatment Effect (ATE), which is the average difference in outcomes between the treated and control groups attributable to the treatment under investigation. However, RCTs have challenges. Attrition,

defined as the departure of participants before the study's conclusion, introduces complexity. Non-random attrition can distort results, compromising the study's reliability.

In the field of advanced methodologies, Principal Stratification emerges as a framework adept at handling post-treatment variables that are influenced by the treatment itself. Principal strata, defined by the causal states that individuals would be in under all treatment conditions, provide a nuanced understanding of treatment effects within specific subgroups. In this project, there are 4 principal strata: Always Post-test (AP) representing the students who would always take the post-test no matter which group they belonged to, Treatment Post-test (TP) representing the students who would only take the post-test if assigned to the treatment group, Control Post-test (CP) representing the students who would only take the post-test if assigned to the control group, and Never Post-test (NP) representing the students who would never take the post-test no matter which group they belonged to.

One of the primary goals is to estimate the ATE for the Always Post-test AP stratum, or estimate the Survivor Average Treatment Effect (Bia et al., 2022). Monotonicity assumption ensures that none of the students belonged to TP - students who were assigned to Control and did not take the post-test would also not take the post-test if assigned to Treatment, and students who were assigned to Treatment and took the post-test would also take the post-test if assigned to Control. Principal Scores, representing the predicted probabilities of belonging to a specific principal stratum, remain essential in addressing potential biases when estimating causal effects, which explain the impact or influence of a particular treatment or intervention, within strata.

The M-estimator is a way to estimate ATE in the AP stratum. It relies on Residualized Covariate Ignorability, assuming that the treatment assignment is independent of the potential outcomes given the observed covariates and their residuals. This assumption posits independence of potential outcomes and pre-treatment covariates conditional on principal scores and treatment assignment. However, estimating principal scores is still a challenging task. To solve this, previous methods include Logistic Regression and Bayes Theorem. However, these methods produced scores outside the meaningful probability range, posing limitations.

Methodology

In this chapter, the project concentrates on estimating the treatment effect of the Always Post-test (AP) stratum, in which the student would always take the post-test, and determining principal scores for AP in the control group. The process begins with a Multinomial Logit Model, utilizing the Expectation-Maximization (E-M) algorithm to categorize students into principal strata (AP, NP, CP). Key functions such as $Estepi$, $Estep$, $Mstepi$, and $Mstep$ are introduced to address missing principal scores through an iterative convergence of the E-M algorithm.

Subsequent steps involve determining treatment effects through the construction of two linear models - one including both observations and principal strata, and the other considering only observations. The Bootstrap method is employed to estimate the standard error of the treatment effect, offering insights into its variability across 500 bootstrap iterations. Additionally, the Bootstrap approach generates a 95% confidence interval, providing a nuanced understanding of plausible values for the population treatment effect. The R codes for all these steps can be found in Appendix.

Data

The Data chapter delves into an educational experiment targeting middle school students to enhance their math skills, employing interventions such as From Here to There (FH2T), Dragon Box 12+ (DragonBox), and ASSISTments. Notably, FH2T and DragonBox utilize game-based approaches, fostering engagement through immersive gameplay and gradually introducing complex algebraic concepts. In contrast, ASSISTments serves as a non-game platform, providing personalized assignments and immediate feedback to support classroom learning. The dataset encompasses 3,271 students in a randomized controlled trial, with a narrowed focus on comparing the impact of game-based technologies FH2T and DragonBox.

Key variables, including post-test outcomes, school identifiers, gender, attendance, pretest scores, race, and post-test scores, among others, contribute to a robust analysis. To address missing data, the "missForest" R package is employed, leveraging a random forest framework for iterative imputation. This comprehensive approach ensures a thorough exploration of the interventions' effectiveness, contributing valuable insights to educational strategies for middle school math improvement.

Results

The results chapter delves into key metrics and models derived from an experimental study. Notably, the attrition rate, indicative of students not completing the post-test, averaged at 40.8%, with varying rates across schools. School #1 exhibited an exceptionally high rate of 91%, emphasizing the influence of institutional engagement on test completion. Higher attrition rates were associated with students with elevated pre-test scores, suggesting a potential correlation between academic proficiency and test completion. Further analysis through principal score models unveiled coefficients illustrating preferences for certain categories over baselines, with school affiliation and pre-test scores prominently influencing student choices.

The investigation extended to treatment effects, comparing the impact of DragonBox and FH2T on post-test math scores. The effect model indicates that on average, students using DragonBox have a 18.5% performance in the post-test than in the pre-test. The observed post-test model reflects a positive treatment effect but at a slightly lower magnitude of 16.9%. Rigorous examination via the Bootstrap method introduced uncertainty but reinforced the robustness of the effect model, suggesting potential improvements of up to 66.9% and 45.3%, or negative effects of up to 29.9% and 11.1%, in post-test scores with DragonBox, depending on whether the Control Post-test (CP) stratum is considered in the model. This comprehensive analysis underscores the nuanced factors influencing attrition rates and substantiates the positive impact of DragonBox on student performance in post-test math scores.

Conclusion

This project addressed challenges in middle school mathematics education exacerbated by COVID-19, focusing on game-based technologies like FH2T and DragonBox. High attrition rates prompted a departure from traditional methods, leading to an innovative approach - multinomial regression with the Expectation-Maximization (E-M) algorithm. This methodology provided a nuanced understanding of student performance and effectively managed missing data, surpassing limitations of previous techniques. However, complexities in implementing the E-M algorithm and Bootstrap's method, along with reliance on assumptions, pose challenges. Future research should explore alternative, concise methods and address questions regarding assumption validity, ensuring broader applicability and reliability in enhancing middle school math education. Overall, this project contributes methodological advancements, refining the analysis of game-based technologies' impact on algebraic comprehension and advancing the reliability of educational research methodologies.

1 Introduction

In recent years, the struggle of many middle school students to grasp mathematical concepts, compounded by the educational disruptions caused by the COVID-19 pandemic, prompted a team of researchers to launch an ambitious experiment. Their primary goal was to address the challenges in learning mathematics, particularly algebra, among middle schoolers. The experiment employed a scientifically rigorous approach, featuring a well-structured design and a meticulous student selection process through randomization to ensure a representative sample. Students were assigned to various groups, each receiving a different educational intervention, with the aim of evaluating the effectiveness of innovative educational technologies in enhancing algebraic understanding. This study seeks to provide valuable insights into potential solutions for improving mathematical education at the middle school level.

The experiment emphasized three educational technologies: From Here to There (FH2T), Dragon Box 12+ (DragonBox), and ASSISTments, chosen for their innovative approaches to teaching algebra and potential to enhance learning. FH2T employed real-world scenarios and problem-solving tasks, DragonBox utilized interactive gameplay, and ASSISTments provided personalized feedback and reports. The study aimed to investigate the impact of these distinct instructional strategies on algebraic comprehension. This project specifically focuses on FH2T and DragonBox, as both employ game-based elements for learning, allowing for a comparative analysis to uncover the most effective aspects of game-based education in enhancing mathematical proficiency among middle school students and optimizing their use in educational settings.

One of the most significant challenges encountered during this experiment was the high attrition rate, the dropout of participants over the study's duration, a particularly pressing concern in the realm of educational technology research. Attrition poses substantial issues, potentially undermining the validity and reliability of findings. In the context of assessing the impact of educational technologies on middle school students' algebra comprehension, attrition threatened the study's integrity. It's problematic primarily because it can introduce sample bias,

as the characteristics of those who drop out may differ significantly from those who complete the study, potentially distorting results. This discrepancy is especially troubling in educational settings, where it may obscure the true effects of interventions being examined.

Researchers have traditionally employed various methods to address attrition, with one common approach being the exclusion of dropouts from the analysis. However, this disrupts the experiment's randomization that ensures each participant has an equal chance of being assigned to any treatment group, potentially leading to biased results. Another conventional method is to assume that dropouts have a post-test score of zero, maintaining the sample size but introducing its own bias by not accurately representing their potential outcomes if they had stayed in the study. These challenges in handling attrition complicate the interpretation of treatment effects and underscore the importance of addressing attrition issues carefully in educational technology research.

In addressing the challenges of educational research, various analytical methods have been employed previously. One method is to use the concept of principal stratification under monotonicity assumption. This approach involves categorizing individuals based on their potential outcomes, assuming a consistent direction of the effect of the treatment on whether the students took the post-test or not. In particular, we can think of estimating the average effect of people who would always take the post-test regardless of randomization. These people are categorized into Always Post-test (AP) stratum, which is sometimes referred to as the concept of Survivor Average Causal Effect. Based on the categories, or strata, the Logistic Regression (LR) model combining with Bayes Theorem is used to estimate probabilities and make inferences about the population. However, this method can sometimes produce probabilities that exceed the logical limit of one, indicating a potential flaw in the model when applied to certain types of data. Another method includes M-estimation, which involves creating estimators that are robust against certain types of modeling misspecifications. When combined with residualized covariate ignorability - a concept that assumes that any covariate related to both the treatment and the outcome is accounted for in the model - it can provide a powerful

tool for analysis. However, this method can be complex to implement and may require stringent assumptions that are not always met in practical educational research scenarios.

In addressing the challenges posed by the high attrition rate in the study, this project introduces an innovative method to address the challenges posed by a high attrition rate in a study, departing significantly from traditional techniques. The core of this novel approach lies in the utilization of multinomial regression in conjunction with the Expectation-Maximization (E-M) algorithm. Multinomial regression is particularly well-suited for handling categorical data with more than two outcomes, crucial in contexts where educational outcomes can vary widely and are not limited to binary outcomes. By employing this regression model, the study can capture the nuances and variances in student performance and engagement more accurately.

The E-M algorithm further enhances this approach by effectively managing missing or incomplete data, a common issue in studies with high attrition rates. Operating in two phases - the Expectation (E) step, which estimates missing data based on observed data, and the Maximization (M) step, which computes maximum likelihood estimations with the newly estimated data points - the E-M algorithm iteratively converges to a stable solution. This provides a robust means of handling missing data without relying on biased assumptions seen in traditional methods.

The integration of the E-M algorithm with multinomial regression proves particularly advantageous in this context, allowing for a more nuanced and statistically sound treatment of attrition issues. This combination addresses limitations such as probability estimates exceeding one, a concern in methods like logistic regression and M-estimation, while offering a refined analysis of diverse educational outcomes. Ultimately, this combined approach excels in handling complex data structures and incomplete datasets, providing a more accurate and comprehensive analysis, crucial in educational research where data complexity and attrition are common challenges.

This report is designed to provide a comprehensive overview of the project and the findings. Chapter 2 delves into the background of previous methods used in similar studies, providing context for the current research. Chapter 3 elaborates on the methods proposed in this project, including multinomial regression and E-M algorithm for estimating the probabilities, M-estimation in OLS model to determine the treatment effect, and Bootstrap method for estimating standard errors. Chapter 4 offers an in-depth look at the data used in the study, ensuring transparency and thorough understanding of the data sources and characteristics. Chapter 5 presents the results and analysis, highlighting the key findings and their implications for educational technology in middle school mathematics education.

2 Background

This chapter discusses the concepts related to the project as well as presents the methods that had been tried previously.

2.1 Causal Inference Framework

The causal inference framework forms the foundation of empirical research, aiming to ascertain the effects of various interventions or treatments under study. This framework, central to both the social and natural sciences, relies heavily on the concept of potential outcomes, a key theoretical approach in understanding the direct impact of treatments.

2.1.1 Potential Outcomes

Let Z represent the group that each student was assigned to. $Z = 1$ if a student is in the treatment group, whereas $Z = 0$ if a student is in the control group.

Let Y represent the post-test score of each student and $Y(Z)$ represent the post-test score for a specific Z. In other words, $Y(1)$ represents the post-test score for students in treatment group, whereas $Y(0)$ represents the post-test score for students in control group (Rubin, 1974). If a student did not take the post-test, then Y would be a missing value. Each student has both $Y(1)$ and $Y(0)$, but only at most one of them is observed.

Let Z_i represent the group that a specific student i belong to, and let Y_i represent the observed post-test score of that student. Thus, for instance, if $Z_i = 0$, then $Y_i = Y_i(0)$, since $Y_i(0)$ represents the post-test score that student i would get if i were assigned to the control group.

At the core of the causal inference framework are two potential outcomes, $Y(0)$ and $Y(1)$. These outcomes represent the possible states of an individual or a system in the presence or absence of a treatment or intervention. As mentioned previously, $Y(1)$ is the outcome observed under the treatment, while $Y(0)$ is the outcome observed under control. A critical aspect of this concept is that for any individual subject, only one of these outcomes is observable - either the state with the treatment or without it (ThinkCausal, n.d.). This limitation forms the "fundamental problem of causal inference (Holland 1986), “posing challenges in directly observing the counterfactual state.

2.1.2 Individual Treatment Effect (ITE)

The Individual Treatment Effect (ITE) is measured as the difference between the two potential outcomes, $Y(1) - Y(0)$. This metric provides insights into the impact of the treatment on an individual level, allowing researchers to understand the direct effect of an intervention on each subject. However, the challenge lies in the fact that nobody can ever observe both outcomes for the same individual simultaneously, making the estimation of this effect complex and often requiring sophisticated statistical methods (Gelman et al., 2021).

2.1.3 Average Treatment Effect (ATE)

The Average Treatment Effect (ATE) is a statistical measure used in the field of causal inference to estimate the overall or average impact of a specific treatment or intervention across a group or population. It calculates the difference between the average outcome of the treated group and the average outcome of the untreated group, providing an aggregate assessment of the treatment's effectiveness on a larger scale. In other words, ATE can be obtained by calculating the average of the ITEs. ATE is valuable for making generalizable conclusions about the effectiveness of an intervention within a specific population, helping policymakers and researchers make informed decisions about the adoption of treatments or policies.

2.2 Randomized Controlled Trials (RCT)

A Randomized Controlled Trial (RCT) is a rigorous and widely recognized research design used in the field of experimental research and clinical studies. In an RCT, participants are randomly assigned to different groups, typically including a treatment group that receives the intervention or treatment being studied and a control group that does not receive the treatment or receives a placebo. Randomization helps ensure that any observed differences in outcomes between the groups are more likely to be attributed to the treatment itself rather than confounding factors.

The strength of RCTs lies in their ability to provide unbiased estimates of treatment effects.

When the assignment to treatment is random, the mean outcome Y for the treatment group reliably estimates the mean of the potential outcome $Y(1)$ for the entire population. Similarly, for the control group, the mean of Y represents an unbiased estimate of $Y(0)$. This methodology allows for a clear comparison of outcomes with and without the treatment (ThinkCausal, n.d.), as well as the unbiased estimate of the ATE.

2.3 Attritions in RCTs

Attrition in an experiment refers to the phenomenon where participants who were initially part of the study or treatment group drop out or discontinue their involvement before the experiment's conclusion. This can occur for various reasons, such as personal circumstances, loss of interest, or adverse reactions to the treatment. The loss of participants can lead to biased results, especially when the characteristics of those who leave the study differ substantially from those who remain.

The impact of attrition on RCTs is multifaceted. Primarily, it can lead to a reduction in sample size, diminishing the statistical power of the study. More critically, if the attrition is not random and correlates with either the treatment or the outcome, it can introduce systematic bias. This bias occurs because the final sample may no longer be representative of the original population, skewing the estimated effects of the treatment (Zhang & Rubin, 2003). In other words, if treatment and control group are comparable after randomization, they may no longer be comparable after attrition, since the subjects may attrit in one condition than in another.

To illustrate the potential impact of attrition on study outcomes, consider the following example involving a scientist examining the effects of a specific diet on weight loss. Assume that there were 10 people assigned to the treatment group consuming diet food during the experiment and 10 other people assigned to the control group consuming regular food during the experiment. The initial weights of all 20 people were measured, and both groups had the same mean and standard deviation. After a month, both groups had 5 people who dropped out, and the weights for the rest of the people were measured. The result is that the average weight in the treatment

group is much lower than the average weight in the control group. However, the fact is that the people who dropped out in the treatment group had higher initial weight than the people who did not drop out, whereas the people who dropped out in the control group had lower initial weight than the people who did not drop out. This discrepancy highlights the potential distortion in results due to non-random attrition, cautioning against simplistic conclusions regarding the diet's efficacy in weight loss. In conclusion, the susceptibility of randomized controlled trials to attrition underscores the importance of diligently addressing and analyzing participant dropout, as it can introduce bias and compromise the reliability of study findings.

2.4 Principal Stratification

2.4.1 Basic Definitions

Let S represent the observation of whether a student took the post-test. $S = 1$ if a student took the post-test, whereas $S = 0$ if a student did not take the post-test.

Among all the students, there are 4 types of different observations in total. The number of each observation should be recorded. Table 1 shows the relationships of the 4 observations.

Table 1: Observations of Students

	$\boldsymbol{Z = 1}$	$\boldsymbol{Z = 0}$
$\boldsymbol{S = 1}$	$S(1) = 1$: students in the treatment group that took the post-test.	$S(0) = 1$: students in the control group that took the post-test.
$\boldsymbol{S = 0}$	$S(1) = 0$: students in the treatment group that did not take the post-test.	$S(0) = 0$: students in the treatment group that did not take the post-test.

Principal stratification is a statistical framework that provides a way to handle causal inference in the presence of post-treatment variables that are affected by the treatment itself. This framework is particularly useful in the context, where the study is interested in the causal effect of a treatment on students' test scores, but not all students took the post-test.

Principal strata are defined by potential outcomes under different treatment conditions. In this project, there are 4 principal strata: Always Post-test (AP), Treatment Post-test (TP), Control

Post-test (CP), and Never Post-test (NP). Table 2 represents the relationships of the 4 strata (Zhang & Rubin, 2003).

Table 2: Principal Stratification of Students

	$S(1) = 1$	$S(1) = 0$
$S(0) = 1$	AP: students that take the post-test no matter which group they belong to.	CP: students that take the post-test if they are assigned to control group and do not take the post-test otherwise.
$S(0) = 0$	TP: students that take the post-test if they are assigned to treatment group and do not take the post-test otherwise.	NP: students that do not take the post-test no matter which group they belong to.

The goal of principal stratification is to estimate the proportion of individuals in each principal stratum, especially the average effect within each stratum. In this project, for everyone with an observed outcome, AP and TP students can be compared in the treatment group, whereas AP and CP students can be compared in the control group. Therefore, the next goal is to estimate the treatment effects for the AP stratum. This is often referred to the Survivor Average Causal Effect.

2.4.2 Monotonicity Assumption

The monotonicity assumption in principal stratification implies that if a student takes the test under the treatment condition, they would also take the test under the control condition. In other words, no students belong to the TP stratum. Therefore, $S(1)$ never becomes larger than $S(0)$. This assumption is crucial for the analysis because it allows to infer the behavior of students under the control condition based on their behavior under the treatment condition. Table 3 shows the principal strata of students based on the observation.

Table 3: Principal Strata Under Monotonicity Assumption

	$Z = 1$	$Z = 0$
$S = 1$	All students are in CP stratum.	Students are in either AP or CP.
$S = 0$	Students are in either CP or NP.	All students are in NP stratum.

However, in this project, the monotonicity assumption can lead to a huge error if a great number of students actually belong to the TP stratum.

2.5 Principal Score

Principal scores are the probabilities of belonging to each principal stratum, given the observed pre-treatment covariates X. These scores can be used to adjust for confounding in the estimation of causal effects within strata. All principal scores can be represented in a formula. For example, the probability of being in the AP stratum conditional on X is $P(S(1) = 1 \; and \; S(0) = 1|X)$.

2.5.1 Principal Ignorability

Principal Ignorability (PI) allows to estimate the causal effect of the treatment on the test scores, independently of the principal stratum to which a student belongs. This is crucial for the analysis because it allows to estimate the missing test scores independently of whether a student took the test or not. There are two types of PIs: weak PI and strong PI. Weak PI assumes that the scores are same if Z=0, and strong PI assumes that the scores are same if $Z = 0$ and $Z = 1$. Here are two formulas representing the concept of PI (Feller et al., 2017):

$$E(Y(0)|X, S(1) = 1) = E(Y(0)|X, S(1) = 0) = E(Y(0)|X) \ldots\ldots (1)$$
$$E(Y(1)|X, S(1) = 1) = E(Y(1)|X, S(1) = 0) = E(Y(1)|X) \ldots\ldots (2)$$

where E represents expected value. Weak PI assumes that formula (1) is true, and strong PI assumes both formulas are true.

However, PI also has disadvantages. In this project, since the attrition rate is high, the missing data may not be random. Therefore, the PI assumption should be avoided, since it may lead to biased estimates.

2.5.2 M-estimator and Residualized Covariate Ignorability

The M-estimator is a method for estimating the principal effects, the average effect within each principal stratum, based on the observed data. It "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". The basic formula is:

$$\sum_{i=1}^{n} \psi_i(Y_i, \theta) = 0$$

where ψ_i represents a function of Y_i and θ, and both of them should be solved without depending on i or n (Stefanski & Boos, 2002).

Residualized covariate ignorability is an assumption that allows to use the M-estimator. It assumes that, conditional on the principal scores and the treatment assignment, the potential outcomes are independent of the pre-treatment covariates. Here is the basic formula (Sales, 2022):

$$E(Y_C - \gamma' X | X, S_T) = E(Y_C - \gamma' X | S_T),$$
$$E(Y_T - \gamma' X | X, S_T) = E(Y_T - \gamma' X | S_T).$$

Among the observation of $S(0) = 1$, students belong to either AP or CP stratum. However, one of these strata always has less attrition. Principal score can be used to determine the number of students in AP stratum among both strata: $p = P(AP |\ AP\ or\ CP, X)$. The effect can be estimated by OLS using the following formula:

$$Y = b0 + b1\ (1 - S) + b2 * Z + b3 * X\ + error$$
$$=> E(Y|Z,X) = b0 + b1 * E(1 - S|Z,X) + b2 * Z + b3 * X$$
$$=> E(Y|Z,X) = b0 + b1 * (1 - p) + b2 * Z + b3 * X\ since\ E(1 - S|Z,X) = 1 - p$$
$$=> Y = b0 + b1 * (1 - p) + b2 * Z + b3 * X + error.$$

The problem of obtaining p should be solved in this process.

2.6 Previous Methods

The methods that have been tried previously include Logistic Regression and Bayes Theorem (Dufresne, 2023).

2.6.1 Logistic Regression (LR)

Logistic regression is often used when there is a binary outcome. It is a statistical method that can be used to estimate the principal scores. It models the log-odds of the probability of belonging to a particular principal stratum as a linear function of the observed pre-treatment covariates. There are two types of LR models that were tried previously. One model estimates each principal score with LR-fit to different samples ($Z = 1$ or $Z = 0$). The other model fits one big LR with interactions between X and Z.

2.6.2 Bayes Theorem

Bayes theorem is a fundamental principle in probability theory and statistics that describes how to update the probability of a hypothesis based on evidence. It can be used to estimate the principal scores by updating the prior probabilities of belonging to a particular principal stratum based on the observed data. The basic formula of Bayes Theorem is: $P(A|B) = P(B|A) * P(A)/P(B)$. Using this formula, it can be concluded that $P(AP|AP\ or\ CP) = P(AP)/P(AP\ or\ CP)$, because $P(AP\ or\ CP|AP) = 1$. Since $P(AP|X)$ can be calculated from the data of treatment group and $P(AP\ or\ CP|X)$ can be calculated from the data of control group, $P(AP|AP\ or\ CP)$ can be obtained by taking their ratio.

2.6.3 Problems with Previous Methods

While these methods have been useful, they have limitations. For example, they may result in a principal score greater than 1, which is not meaningful as a probability. In the previous study, the result shows that the maximum $P(AP|AP\ or\ CP)$ is 1.16 instead of between 0 and 1. Furthermore, they may not adequately adjust for confounding, leading to biased estimates of the causal effects. New methods that address these issues are needed to improve the

understanding of the treatment's effectiveness. These methods will be discussed in detail in the next chapter.

3 Methodology

From Chapter 2, it is discovered that students are in one of the three principal strata: Always Post-test (AP), Never Post-test (NP), or Control Post-test (CP). The goal of this chapter is to estimate the effect of AP. It is already known that students in treatment group who took the post-test are AP. However, principal strata of the control group are unknown - the students who took the post-test are either AP or CP. Therefore, it is needed to estimate the principal score of AP in the control group.

To solve the problem, a multinomial logit model is built to determine which principal strata the students belong to. All the unknown principal strata are treated as “missing” or hidden. Then, the model is fitted with E-M algorithm. The algorithm is repeated several times until the result converges.

Besides this, the project is also interested in the treatment effect, as well as the standard error. To find out the treatment effect, two linear models are built, one with observations and the principal strata, and the other with observations only. Then, the method of Bootstrap is used to determine the standard error. All of these steps were performed in R codes (R Code Team, 2023), which can be found in Appendix.

3.1 Multinomial logistic regression

Multinomial regression is an extension of logistic regression that is used when the dependent variable has more than two categories. In this project, the dependent variable is the principal stratum, which has three categories: AP, NP, and CP. The multinomial regression model estimates the probabilities of each category as a function of independent variables. It is a useful method for predicting categorical outcomes when dealing with more than two outcome options. The multinomial regression model works by modeling the log-odds of each category relative to the baseline category, which can be randomly chosen among the outcome categories. In this project, CP is chosen to be the baseline category. The model uses a set of predictor variables to estimate these log-odds and then applies the softmax function, a function that converts a vector of real numbers to a probability distribution, to convert these log-odds into probabilities for each category. The model produces coefficients for each predictor variable, which indicate how strongly they influence the likelihood of belonging to a particular stratum.

The formulas of the model would be:

$log(P(AP)/P(CP)) = \theta_{AP}^{T} X,$ and
$log(P(NP)/P(CP)) = \theta_{NP}^{T} X,$

where θ represents the coefficients, and X represents the covariates (PennState, n.d.).

Since the strata cannot be observed for everyone, E-M algorithm is needed to determine the principal strata and their principal scores. The details of this algorithm will be discussed in the next section.

3.2 E-M Algorithm

The Expectation-Maximization (E&M) algorithm is a general iterative optimization method used to estimate the parameters of statistical models, particularly when dealing with incomplete or missing data. It's a two-step process that alternates between the E-step (Expectation step) and the M-step (Maximization step) to iteratively refine parameter estimates.

In the E-step, the primary goal is to calculate the expected values of the missing data or latent variables given the current parameter estimates. This step involves computing the posterior probabilities or responsibilities associated with the latent variables. These responsibilities represent the likelihood that a particular observation belongs to a particular category or class. In the M-step, the primary goal is to update the parameter estimates to maximize the expected log-likelihood calculated in the E-step. This step involves finding the parameter values that optimize the expected log-likelihood, treating the expected values of the latent variables as if they were observed.

The algorithm is appropriate in this project because the complete data likelihood is intractable. Not all the information needed to fit in with the model are present. Therefore, it is needed to use this algorithm to fit the missing data.

3.2.1 E-M Algorithm for Principal Scores

Table 4 shows the formulas of the conditional probabilities of latent strata for each of the 4 groups (Ding & Lu, 2016).

Table 4: Conditional Probabilities of Latent Strata

	$\boldsymbol{Z=1}$	$\boldsymbol{Z=0}$
$\boldsymbol{S=1}$	$P(U=AP\|X)=1$ $P(U=CP\|X)=0$ $P(U=NP\|X)=0$	$P(U=AP\|X)=1/(1+e^{-\theta_{AP}^{T}X})$ $P(U=CP\|X)=1/(1+e^{\theta_{AP}^{T}X})$ $P(U=NP\|X)=0$
$\boldsymbol{S=0}$	$P(U=AP\|X)=0$ $P(U=CP\|X)=1/(1+e^{\theta_{NP}^{T}X})$ $P(U=NP\|X)=1/(1+e^{-\theta_{NP}^{T}X})$	$P(U=AP\|X)=0$ $P(U=CP\|X)=0$ $P(U=NP\|X)=1$

Under the monotonicity assumption, there might be some students who has $S=0$ while $Z=1$ and $S=1$ while $Z=0$, but there are no students who has $S=0$ while $Z=0$ and $S=1$ while $Z=1$. Therefore, if $S=1$ and $Z=1$, then the student must be AP, and if $S=0$ and $Z=0$, then the student must be NP.

If $S = 1$ and $Z = 0$, then the student is either AP or CP. From the model $log(P(AP)/P(CP)) = \theta_{AP}^T X$, it can be obtained that $P(AP) = P(CP) * e^{\theta_{AP}^T X}$. Since AP and CP are disjoint, then $P(AP) + P(CP) = 1$. Therefore, $P(AP) = 1 - P(CP) = P(CP) * e^{\theta_{AP}^T X}$ => $(1 + e^{\theta_{AP}^T X}) * P(CP) = 1$ => $P(CP) = 1/(1 + e\text{^}(\theta_{AP}^T X))$, and $P(AP) = 1 - P(CP) = 1 - 1/(1 + e^{\theta_{AP}^T X}) = 1/(1 + e^{-\theta_{AP}^T X})$. Similarly, if a student has $S = 0$ and $Z = 1$, then the student is either NP or CP, and $P(CP) = 1/(1 + e^{\theta_{NP}^T X})$ and $P(NP) = 1/(1 + e^{-\theta_{NP}^T X})$.

Next, to perform the E-M algorithm, here are the key functions defined:

Estepi: This function calculates the E-step for each student (in other word, i). It takes inputs including as Z (0 or 1), indicating treatment group or control group, $hasPosttest$ (0 or 1), indicating whether the student took the post-test, X, representing predictor variables, and $coefs$, which are initially random coefficients. The inputs Z and $hasPosttest$ indicate the group that the student belongs to, and then the function calculates and returns the probabilities of belonging to the "AP", "NP", and "CP" strata for the given student based on the probability formulas of that group.

Estep: This function is responsible for performing the E-step for all students. It takes inputs Z, $hasPosttest$, X, and the coefficients ($coefs$) for all students. It uses the $Estepi$ function for each student and returns a probability table with the probabilities for all students.

Mstepi: This function begins the M-step by creating the weight table for each student (i). It takes inputs such as i, Z, $hasPosttest$, and the probabilities ($probs$) calculated in the E-step. Within this function, the weight table is created, including i, U, which is either "AP", "CP", or "NP", and w, the weight based on the values in the probability table. The table has either 1 or 2 rows. For the 1-row tables, U is either “AP” or “NP”, and w is always 1. For the 2-row tables, the values of i are the same for both rows. The first row of U is either “AP” or “NP”, and the second row of U is always “CP”. The values of w are equal to the principal scores

calculated from the E-step.

$\boldsymbol{Mstep}$: This function manages the M-step for all students. It takes inputs Z, $hasPosttest$, X, and the probability table ($probs$) for all students. In this function, weight tables for all students are combined, and then a multinomial regression model is fitted using "CP" as the baseline category. The model is a weighted multinomial logit, with weights equal to estimated probabilities. The coefficients from this model are extracted and used as the updated coefficients.

To see why the model makes sense, it can be first assumed that there is an infinite number of people in a population. Assuming that the principal scores are correct, each person in the sample represents all people with the same covariates. For example, if there is a white male from school #1, and he is 90% AP and 10% CP, then among all white males from school #1 in the population, there are 90% of people in AP and 10% of people in CP. The data from this experiment can be regarded as a finite random sample from the population, and in the E-M algorithm, the samples are weighted to represent the population.

3.2.2 Iteration step

The iteration step alternately performs E-step and M-step until the convergence is achieved. To execute this step, random coefficients ($coefs$) are initialized as the initial coefficient table, and then these coefficients are used along with the $Estep$ function to calculate the initial probability table ($probs$). The previous-probability table ($probs1$) is initialized with values set to 1 plus the initial probability table. Then, a series of E-steps and M-steps is performed until a convergence criterion is met. In this project, the convergence criterion is that all values in the two probability tables have a difference below the tolerance, which is set to be 0.001. In each iteration, the new probability table is calculated using the M-step and E-step. The process is repeated until the convergence criterion is met.

3.3 Determining Treatment Effect

As mentioned in sec 2.6.2, the M-estimation in OLS model involves minimizing a certain objective function to estimate the parameters of the model. In the context of estimating treatment effects, the OLS model is often applied to analyze experimental data with observed post-test scores. The covariates included in the model typically consist of various factors that may influence the outcome of interest, such as pretest scores, demographic characteristics, or other relevant variables. The crucial aspect of the OLS model in estimating the treatment effect lies in the coefficient associated with the treatment variable. This coefficient quantifies the difference in the outcome variable between the treatment and control groups, thereby providing a direct measure of the treatment's impact while controlling for the effects of the included covariates. By examining this coefficient, the effectiveness of the treatment can be accessed while accounting for potential confounding factors, making OLS a valuable tool for causal inference in experimental studies.

The treatment effect, or ATE, represents the average difference in outcomes between the treatment group and the control group. In this project, the treatment effect specifically measures how the treatment affects post-test scores. A positive treatment effect indicates that the treatment has a beneficial impact on students' performance, while a negative effect suggests the opposite.

To determine the ATE, an OLS model, called “effect model”, is built to estimate the ATE for the AP stratum. Before building the model, the data for students who took the post-test is extracted from the original dataset. The model is built based on the Z (0 or 1), covariates X and the probability variable (P) obtained from the $probs$ table. Then, as a comparison, a second OLS model, called “observed post-test model”, is fit. This model only includes Z and the covariates X and does not include the principal scores. This is the current conventional approach that assumes that attrition is ignorable, or there is no CP stratum. In both models, emphasis is placed on the coefficients of Z, as they signify the Average Treatment Effect (ATE) of the treatment. It is worth noting that the Effect Estimate is derived from linear regression

involving principal scores, while the regression without principal scores yields the conventional effect estimate, assuming a binary classification of students as either AP or NP. This dual approach allows for a comparison between the method and the conventional one to assess if the method yields different results.

3.4 Bootstrap

The primary goal of bootstrapping is to estimate the standard error of the treatment effect, which helps gauging the variability of the estimate. Additionally, the bootstrap samples will be used to calculate a 95% confidence interval for the treatment effect, which provides a range of plausible values for the population treatment effect.

To perform the bootstrap, the sample size is set to be the total number of students (n), and the number of bootstrap iterations (B) is set to be 500. For each of the 500 bootstrap iterations (b), a random sample from the original dataset is taken as a replacement. Within each bootstrap iteration, the prognostic scores are estimated using the E-M algorithm, similar to the procedure outlined in earlier sections of this chapter. This involves updating the coefficients and probabilities using the E-step and M-step until convergence is achieved. After estimating the prognostic scores, the treatment effect is estimated in each bootstrap sample. A subset of the sample consisting of students who took the post-test is created, and then the probability of being in the "Control Post-test (CP)" stratum (P) is calculated. Two models are fitted: the "effect" model and the "observed post-test" model. The treatment effects are extracted from these models. These steps are repeated for all 500 bootstrap iterations, generating 500 treatment effect estimates for both the "effect" model and the "observed post-test" model.

The standard error of the treatment effect is calculated as the standard deviation of the treatment effect estimates obtained from the bootstrap samples. It provides a measure of the variability of the treatment effect estimate across different samples, helping people understand the precision of the estimate.

The 95% confidence intervals for both effect model and observed post-test model are calculated using the ATE calculated previously plus and minus the 1.96 times the standard error obtained from the bootstrap. The confidence interval provides a range of values in which it can reasonably be expected the true treatment effect to fall with 95% confidence.

4 Data

This chapter describes the data used for this project.

4.1 Interventions of the Experiment

The data used in the project, also known as fh2t data, is based on an educational experiment on middle school students. The purpose of this experiment was to find out ways to help the students study math better. In this experiment, 3 educational technologies are involved to be interventions: From Here to There, Dragon Box 12+, and ASSISTments. The first two technologies have game-based practices, while the other only has regular practices.

"From Here to There" is designed as an interactive digital game that aims to enhance students' understanding of algebraic concepts through engaging gameplay that incorporates real-world scenarios and problem-solving tasks. The game's structure encourages students to apply algebraic thinking to navigate through various levels, each progressively challenging their skills and understanding. This immersive approach is intended to make learning algebra both accessible and enjoyable, promoting deeper engagement with the subject matter.

"Dragon Box 12+" is an advanced version of the Dragon Box series, tailored for older students with its focus on more complex algebraic concepts. The game uses a unique approach to teaching algebra by gradually introducing abstract symbols and equations through playful, intuitive puzzles. Initially, students interact with objects and characters within the game, which later morph into algebraic representations. This gradual transition from concrete to abstract helps students build a strong conceptual understanding of algebra, making the subject more approachable and less intimidating.

"ASSISTments" is a versatile online platform featured in the study, designed to support and enhance classroom learning. Unlike the other two technologies, ASSISTments is not a game but a tool that provides personalized homework assignments, immediate feedback, and detailed reports to both students and teachers. It allows educators to track student progress in real-time, enabling targeted interventions and support. This platform is particularly useful for identifying areas where students struggle, allowing for timely and effective remediation. ASSISTments' adaptability and data-driven approach make it an invaluable resource in diverse educational

settings, particularly during the challenges posed by the COVID-19 pandemic. There are 2 versions of ASSISTments in the study, one with immediate feedback and hints during the practices, and the other is an "active control" condition where there are no hints available, and the feedback can only be viewed after finishing the practices.

4.2 Data Overview

To test the effectiveness of these interventions, the experiment measured the pre-test scores of the students before using the technologies and their post-test scores after using the technologies.

The full dataset consists of data for 3,271 students that belong to one of the 4 experimental groups: FH2T, DragonBox, Instant, and Delay. The experiment involves a randomized controlled trial (RCT), which means that all the students are randomly assigned to the 4 groups. Since the data comes from an RCT, the randomization assumption is satisfied, which means that the treatment assignment process is unbiased and independent of any potential confounding variables, thus enhancing the internal validity of the study and allowing for more reliable causal inferences to be drawn from the observed outcomes. Among the 4 groups, the “FH2T” and “DragonBox” groups represent the students using From Here to There and Dragon Box 12+ respectively. The students in “Instant” and “Delay” both used ASSISTments as their interventions, where students in “Instant” could get immediate feedback during the practices, while the students in “Delay” could not. Therefore, it is appropriate to compare FH2T and DragonBox, since both groups used a game-based technology, or Instant and Delay, since both groups used the same technology with different form of practices.

In this project, it is believed that game-based technology has stronger effect than regular practice technology, so FH2T VS DragonBox is chosen to be analyzed in the project instead of Instant VS Delay. After this selection, there are 1,960 students in total to be analyzed in this project. Since DragonBox requires the students to download an app while FH2T does not, it is believed that the DragonBox group has higher attrition rate than the FH2T group, and there exists students who would not be able to take the post-test if assigned to DragonBox but would

be able to take the post-test if assigned to FH2T, since some students may have trouble downloading the app. However, it can be assumed that a student in DragonBox who took the post-test would also take the post-test in FH2T, and a student in FH2T who did not take the post-test would also not take the post-test in DragonBox. Therefore, DragonBox is assigned to be the treatment group, and FH2T is assigned to be the control group, since monotonicity assumption can be held under this condition.

4.3 Variables

Each student has 230 variables in total. Among all variables, the study is interested in randomized conditions, the outcomes, and the covariates that may affect the outcome. Table 5 describes all variables that are kept in the analysis.

Table 5: Table of Variables

Variable	Type (C for categorical and N for numerical)	Role	Description	Interpretation
Posttest	Binary	Outcome	1 for students who took the post-test, and 0 for students who did not take the post-test	1 - 59.2% 0 - 40.8%
SchIDPre	C	Covariate	The school that a student attended	1 - 8.5% 2 - 4.1% 3 - 6.4% 4 - 10.1% 5 - 10.1% 6 - 40.1% 8 - 6.8% 9 - 6.8% 11 - 6.2%
condition_ assignment	C	Randomized condition	Treatment: DragonBox Control: FH2T	DB - 33.4% FH - 66.6%
Gender	C	Covariate	M for male and F for female	M - 52.6% F - 47.4%
Present.Days7	N	Covariate	Number of hours	Range: 0-175

			that a student presented in every 7 days	Mean: 164.1 SD: 18.9
pre.total_math_score	N	Covariate	Pretest score of a student	Range: 0-10 Mean: 4.18 SD: 2.53
pre.avg_time_on_tasks	N	Covariate	Average number of hours that a student worked on the tasks	Range: 0-1020 Mean: 84.1 SD: 209.3
accelerated	Binary	Covariate	1 for students who took the accelerated class, 0 otherwise	1 - 19.4% 0 - 80.6%
race	C	Covariate	The race of a student	Asian - 26.0% Black - 5.1% Hispanic - 15.3% White - 49.5% Other - 4.1%
post.total_math_score	N	Outcome	Post-test score of a student	Range: 0-10 Mean: 3.89 SD: 2.93

The outcome is whether a student took the post-test as well as the post-test score. The randomized conditions are treatment (DragonBox) and control (FH2T). There are 7 covariates, including 3 categorical variables, 3 numerical variables, and 1 binary variable. There are a total of 1,960 students from 9 different schools.

4.4 Fitting missing variables

There are some students who have missing values in one or more covariates. To fit the missing data, an R package called “missForest” is used. It operates based on a random forest framework, using an ensemble of decision trees to predict and impute missing values iteratively. This package simplifies the imputation process, making it a valuable tool for researchers and data analysts when working with datasets plagued by missing values, ensuring that meaningful and accurate imputations are made to facilitate subsequent statistical analyses.

5 Results

This chapter displays the results for the project.

5.1 Attrition

The attrition rate is a metric used to measure how many testees are lost during a particular experiment. In this project, the attrition rate represents how many students in the study did not take the post-test. The way to calculate the attrition rate is easy: simply divide the number of students who did not take the post-test by total number of students.

The attrition rates for each randomized condition are: $284/654 = 0.434$ for treatment group, and $516/1306 = 0.395$ for control group. The overall attrition is: $800/1960 = 0.408$.
To find out which factors affect the attrition rate the most, it is needed to get the attrition rates for each covariate. Table 6 shows the attrition rates for each factor.

Table 6: Attrition Rates by Factor

X	**Count of *Posttest*==0**	**Total Count**	**Attrition Rate**
SchIDPre1	152	167	0.910
SchIDPre2	40	81	0.494
SchIDPre3	27	126	0.214
SchIDPre4	32	197	0.162
SchIDPre5	19	198	0.096
SchIDPre6	397	802	0.495
SchIDPre8	51	134	0.381
SchIDPre9	24	133	0.180
SchIDPre11	58	122	0.475
GenderM	428	1030	0.416
GenderF	372	930	0.4
***Present.Days7* >=165**	612	1503	0.407
***Present.Days7* < 165**	188	457	0.411
***pre.total_math_score* >=5**	383	946	0.405
***pre.total_math_score* <5**	417	1014	0.411
***pre.avg_time_on_tasks* >= 110**	288	543	0.530
***pre.avg_time_on_tasks* < 110**	512	1417	0.361
accelerated	114	380	0.3
Not *accelerated*	686	1580	0.434
raceAsian	228	509	0.448
raceBlack	44	99	0.444

raceHispanic	110	300	0.367
raceWhite	384	971	0.395
raceOther	34	81	0.420

It is obvious that the factor that affects the most is the school that the students attended. One of the schools, school #1, has an extremely high attrition rate of over 0.9, whereas some schools, like school #5, have very low attrition rates. This might be because some schools did not pay so much attention to the experiment, so that students may forget about the post-test, whereas some schools paid a lot of attention to the experiment, so that the teachers might urge all the students to complete the post-test.

Another factor that affects the attrition rate might be the pre-test score. Students who had higher pre-test scores had higher attrition rates. This might be because students who had lower grades wanted to practice more, so they regarded the post-test as the final practice, while the students who had higher grades might not need the final practice.

5.2 Coefficients from Principal Score Model

Table 7 shows the coefficients obtained by using the E-M algorithm. In the model, CP is set to be the baseline, and the coefficients represent the relationship between the independent variables and the log-odds of being in a particular category (AP or NP) compared to the baseline category (CP). The "*virtual:accelerated*" represents the students who took both virtual and accelerated class. There are no standard errors in this table, because the model is not the standard multinomial regression model.

Table 7: The Coefficients Table

	AP	**NP**
(Intercept)	2.737	11.636
SchIDPre2	3.640	0.965
SchIDPre3	-7.274	-11.627
SchIDPre4	-6.364	-10.790
SchIDPre5	9.273	3.943
SchIDPre6	-6.517	-9.048
SchIDPre8	-7.658	-10.917

SchIDPre9	6.110	1.792
SchIDPre11	-9.460	-12.042
GenderM	-4.225	-4.065
Present.Days7	0.059	0.024
pre.total_math_score	-0.011	-0.064
pre.avg_time_on_tasks	-0.007	-0.004
accelerated	0.205	-1.341
raceBlack	1.722	1.071
raceHispanic	17.282	16.771
raceother	13.284	12.791
racewhite	1.632	1.265
virtual:accelerated	15.586	16.229

When a coefficient is positive, it shows that as the independent variable increases, the category is preferred over the baseline category, and vice versa. For example, let's compare AP with CP. The factor "*SchIDPre2*" has a positive coefficient, which means that students in school #2 prefer more AP than CP compared to the students from other schools. Another example is "*pre.total_math_score*", which has a negative coefficient. This shows that if a student has a higher pre-test score, it is less likely for the student to be AP than CP. A larger coefficient represents a stronger preference of the category over the baseline category, and vice versa.

For each categorical factor, there is also a "baseline factor". The baseline factors are *SchIDPre1*, *GenderF*, and *raceAsian*. The "Intercept" represents the student who belongs to the baseline factors and has a value of 0 for all numerical factors. In this model, the "Intercept" represents a student who is belonged to School #1, whose gender is Female, who never presents in school, who has a pre-test score of 0, who does not spend any time on tasks, who is not taking the accelerated class, and whose race is Asian. The intercept value is positive for both AP and NP and the value of NP is higher than AP, representing that this student is most likely to be NP and least likely to be CP.

5.3 Coefficients & SEs from effect models

There are two models built for estimating treatment effect: one is effect model, where the post-test score is the outcome, and the covariates and the probability of being CP are independent

variables; the other one is observed post-test model, where the probability of being CP is not included in the model. The main difference between the two models is that the observed model ignores bias, since it assumes that CP=0.

The coefficient of Z represents the treatment effect, or in other words, how is the treatment affecting the post-test score. The effect model has a coefficient of 0.185. This means that, on average, students who were using Dragon Box 12+ could get 18.5% more questions correct on the post-test math scores compared to students using From Here to There, holding other factors constant. This positive coefficient suggests that the treatment had a statistically significant positive impact on math scores. The observed post-test model has a coefficient of 0.169. This coefficient also indicates a positive treatment effect, suggesting that students who were using Dragon Box 12+ could get 16.9% more questions correct on the post-test math scores compared to students using From Here to There. The higher value in the effect model suggests that if assumed that CP students exist, there will be a higher treatment effect.

5.4 Effects estimates, SEs, and CIs

Using the Bootstrap method for $B = 500$ and $tol = 0.001$, the new effect models have a standard error of 0.247, and the new observed post-test models have a standard error of 0.144. The 95% confidence intervals are (-0.299, 0.669) for the effect models and (-0.111, 0.453) for the observed post-test models. This suggests that if considering the principal strata for the students, then it is possible for them to perform 66.9% better using DragonBox, or 29.9% better using FH2T, or anything between. If not considering the principal strata, then it is still possible for students to perform 45.3% better using DragonBox, or 11.1% better using FH2T, or anything between. The result indicates that the effect model has more uncertainty but is more warranted and more robust.

6 Conclusion

This project embarked on a mission to address the pressing challenges in middle school mathematics education, exacerbated by the COVID-19 pandemic, through a meticulous and scientifically rigorous experiment. By focusing on the effectiveness of game-based educational technologies, particularly From Here to There (FH2T) and DragonBox, the study aimed to contribute valuable insights into improving algebraic comprehension among middle school students. The challenges encountered, notably the high attrition rate, prompted a critical examination of traditional methods and the development of an alternative approach. Attrition, a pervasive issue in educational research, has the potential to distort results, introducing bias and compromising the validity of findings.

Examining the landscape of previous methods sheds light on the nuanced nature of addressing attrition. Principal stratification under monotonicity assumption, Logistic Regression combined with Bayes Theorem, and M-estimation with residualized covariate ignorability were among the methods utilized previously. While these approaches presented valuable insights, they also posed challenges, including potential flaws in probability estimates and the complexity of implementation. In response to these challenges, this project introduced an innovative approach that deviates significantly from traditional techniques. Leveraging multinomial regression alongside the E-M algorithm, the aim was to provide a more robust and accurate analysis of the impact of game-based educational technologies while addressing the intricacies of high attrition rates.

Multinomial regression proved instrumental in handling the categorical nature of educational outcomes, offering a nuanced understanding of student performance and engagement. The integration of the E-M algorithm emerged as a powerful tool for managing missing or incomplete data, a prevalent issue in studies with high attrition rates. By iteratively estimating missing data in the Expectation step and refining maximum likelihood estimations in the Maximization step, the E-M algorithm provided a stable and unbiased solution. This combined approach surpasses in handling complex data structures and incomplete datasets, addressing

limitations seen in previous methods. By presenting a statistically sound treatment of attrition issues, the methodology employed in this project contributes to a more accurate and comprehensive analysis - an imperative in the field of educational research where data complexity and attrition are commonplace challenges.

However, there are still problems existing in the methodology. While the E-M algorithm is powerful, it relies on certain assumptions, and its effectiveness can be influenced by the quality of these assumptions. In practical scenarios, meeting all the stringent assumptions of the E-M algorithm may be challenging, potentially affecting the accuracy of the results. Furthermore, the integration of multinomial regression with the E-M algorithm, as well as the Bootstrap's method for estimating the standard error, can be complex to implement. Therefore, it is a worth thinking question that whether there is a more concise and straightforward method to get the same purpose.

Besides that, there are still several questions that future researchers should think about. First, it is important to highlight that the methods in this project are applied based on the monotonicity assumption. When this assumption is not true, the methods may lead to more biased results. Therefore, the ways to determine whether the assumption is true, as well as the methods applied when the assumption is not true, should be deeply considered and developed by future researchers. Moreover, further researchers should also investigate the conditions to apply the methods in this project, as well as the conditions to use other methods, including conventional methods and the methods proposed in the future.

In essence, this project strives to offer a methodological advancement in addressing the challenges posed by attrition in educational technology research. By refining the understanding of the impact of game-based educational technologies on algebraic comprehension, this study contributes to the ongoing dialogue on enhancing middle school mathematics education, paving the way for more robust and reliable research methodologies in the future.

References

Bia, M., Mattei, A., & Mercatanti, A. (2022). Assessing Causal Effects in a longitudinal observational study with "truncated" outcomes due to unemployment and nonignorable missing data. *Journal of Business & Economic Statistics*, *40*(2), 718-729.

Ding, P., & Lu, J. (2016). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(3), 757-777.

Dufresne, L. (2023). An M-Estimator for the Survivor Average Treatment Effect.

Feller, A., Mealli, F., & Miratrix, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics*, *42*(6), 726-758.

Gelman, A., Hill, J., & Vehtari, A. (2021). In *Regression and other stories*. essay, Cambridge University Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rubin, D. B. (1974). Estimating causal effects of Treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*, 688-701.

Sales, A. (2022). GEEPERS: Principal Stratification Using Principal Scores and Stacked Estimating Equations.

Stefanski, L. A., & Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, *56*(1), 29-38.

The Pennsylvania State University. (n.d.). *8: Multinomial Logistic Regression Models: Stat 504*. PennState: Statistics Online Courses.

ThinkCausal. (n.d.). *Causal estimands*. https://apsta.shinyapps.io/thinkCausal/

Zhang, J. L., & Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death." *Journal of Educational and Behavioral Statistics*, *28*(4), 353-368.

Appendix: R codes for Methodology

```
library(nnet)
library(foreign)
library(ggplot2)
library(reshape2)
library(missForest)
library(parallel)

dat = read.csv('fh2tDat.csv')

model_data = dat[, c("Posttest", "SchIDPre", "RACE", "virtual", "courseName",
"condition_assignment", "Gender", "Present.Days7", "pre.total_math_score",
"pre.avg_time_on_tasks")]
model_data$accelerated <- model_data$courseName=="Accelerated Mathematics" |
model_data$courseName=="Accelerated Math - Virtual"
model_data$accelerated <- as.integer(model_data$accelerated)
model_data$courseName <- NULL
model_data$race=c("Hispanic", "AmeInd", "Asian", "Black", "PacIs", "white", "multi")
[model_data$RACE]
model_data$race[model_data$race%in%c('AmeInd', 'multi', 'PacIs')] <- "other"
model_data$SchIDPre <- as.factor(model_data$SchIDPre)
model_data$condition_assignment <- as.factor(model_data$condition_assignment)
model_data$RACE <- NULL
model_data$Gender <- as.factor(model_data$Gender)
model_data$race <- as.factor(model_data$race)

fullData=result$ximp

dotProduct <- function(x,y){
   product=x*y
   dotProduct=sum(product)
   return(dotProduct)
}

DBFHdata=subset(fullData,condition_assignment%in%c("DragonBox", “FH2T"))
DBFHdata$Z <- ifelse(DBFHdata$condition_assignment=="DragonBox",1,0)
X <- model.matrix(Posttest ~ . +virtual:accelerated-virtual -Z -condition_assignment,data=DBFHdata)
coefs=cbind(AP=rnorm(ncol(X)),NP=rnorm(ncol(X)))

Estepi <- function(Z,hasPosttest,X,coefs){

   probs=c(AP=0,NP=0,CP=0)
```

```
  if(Z==1 & hasPosttest){
    probs["AP"]=1
  }
  if(Z==0 & hasPosttest){
    probs["AP"]=1/(1+exp(-dotProduct(coefs[, “AP"],X)))
    probs["CP"]=1/(1+exp(dotProduct(coefs[, “AP"],X)))
  }
  if(Z==1 & !hasPosttest){
    probs["NP"]=1/(1+exp(-dotProduct(coefs[, “NP"],X)))
    probs["CP"]=1/(1+exp(dotProduct(coefs[, “NP"],X)))
  }
  if(Z==0 & !hasPosttest){
    probs["NP"]=1
  }
  return(probs)
}


Estep <- function(Z, hasPosttest,X, coefs){
  probs=matrix(nrow=length(Z),ncol=3)
  colnames(probs)=c('AP', ‘NP', ‘CP')
  for(i in 1:length(Z))
    probs[i,]=Estepi(Z[i],hasPosttest[i],X[i,],coefs)
  return(probs)
}

probs=Estep(DBFHdata$Z,DBFHdata$Posttest,X,coefs)

makeMultDat <- function(dat){
  dat$i=1:nrow(dat)
  dat$U=character(nrow(dat))
  sdat <- split(dat,list(Z=dat$Z,hasPosttest=dat$Posttest))

  sdat <- lapply(sdat, \(x){
    if(x$Z[1]==1 & x$Posttest[1]==1){
      x$U <- 'AP'
    } else if(x$Z[1]==0 & x$Posttest[1]==1){
      x <- rbind(
        within(x,U <- 'AP'),
        within(x,U <- 'CP'))
    } else if(x$Z[1]==1 & x$Posttest[1]==0){
      x <- rbind(
        within(x,U <- 'NP'),
        within(x,U <- 'CP'))
```

```
    } else x$U <- 'NP'
    return(x)
  })
  return(do.call('rbind',sdat))
}


Mstepi <- function(i,Z,hasPosttest,probs){
  if(Z==1 & hasPosttest){
    return(data.frame(i=i,U='AP',w=1))
  }
  if(Z==0 & hasPosttest){
    return(data.frame(i=i,U=c('AP', 'CP'),w=probs[c('AP', 'CP')]))
  }
  if(Z==1 & !hasPosttest){
    return(data.frame(i=i,U=c('NP', 'CP'),w=probs[c('NP', 'CP')]))
  }
  if(Z==0 & !hasPosttest){
    return(data.frame(i=i,U='NP',w=1))
  }
}

Mstep0 <- function(Z,hasPosttest,X,probs){
  multDat=Mstepi(i=1,Z=Z[1],hasPosttest=hasPosttest[1],probs=probs[1,])
  for(i in 2:length(Z))
    multDat=rbind(multDat,Mstepi(i=i,Z=Z[i],hasPosttest=hasPosttest[i],probs=probs[i,]))


  Xmult=X[multDat$i,-1]
  multDat$U <- as.factor(multDat$U)

  multDat$U2 <- relevel(multDat$U, ref="CP")
  test <- multinom(U2 ~ Xmult, data = multDat, weights=w)

  coefs<- t(coef(test))
  return(coefs)
}

Mstep <- function(multDat,Xmult,probs){
  multDat$w <- 0
  for(uu in c('AP', 'CP', 'NP')){
    ind <- which(multDat$U==uu)
    multDat$w[ind] <- probs[multDat$i[ind],uu]
  }
```

```
  multDat$U <- as.factor(multDat$U)

  multDat$U2 <- relevel(multDat$U, ref="CP")
  test <- multinom(U2 ~ Xmult, data = multDat, weights=w)

  coefs<- t(coef(test))
  return(coefs)
}

coefs= cbind(
  AP=coef(glm(DBFHdata$Posttest[DBFHdata$Z==1]~X[DBFHdata$Z==1,-1],family=binomial)),
  NP=coef(glm(1-DBFHdata$Posttest[DBFHdata$Z==0]~X[DBFHdata$Z==0,-1],family=binomial)))

probs=Estep(DBFHdata$Z,DBFHdata$Posttest,X,coefs)
probs1=probs+1

multDat <- makeMultDat(DBFHdata)
Xmult <- X[multDat$i,-1]

iter=0
maxDiff=NULL
while(!all(abs(probs1 - probs) < 0.001)){
  iter=iter+1
  probs2=probs
  invisible(capture.output(coefs<-Mstep(multDat,Xmult,probs)))
  probs=Estep(DBFHdata$Z,DBFHdata$Posttest,X,coefs)
  probs1=probs2
  maxDiff=c(maxDiff,max(abs(probs1-probs)))
  print(paste(iter,round(maxDiff[iter],3)))
  if(iter>500) break
}

mainCoefs <- coefs
mainProbs <- probs

effData=subset(DBFHdata,Posttest==1)
effData$P=probs[DBFHdata$Posttest==1, 'CP']
effData$Y=dat[rownames(effData), 'post.total_math_score']
effMod=lm(Y~.+virtual:accelerated-Posttest-virtual-condition_assignment,data=effData)
obsPostMod=lm(Y~.+virtual:accelerated-Posttest-virtual-condition_assignment-P,data=effData)
eff.est=coef(effMod)['Z']
obs.est=coef(obsPostMod)['Z']
```

```
DBFHdata$Y <- dat[rownames(DBFHdata), 'post.total_math_score']
n <- nrow(DBFHdata)
begin=Sys.time()
B=500
tol = 0.001

est1= function(b) {
   cat('.')
   rows <- sample(1:n,n,replace=TRUE)
   newdat=DBFHdata[rows,]
   newX <- X[rows,]

   multDat <- makeMultDat(newdat)
   Xmult <- newX[multDat$i,-1]
   coefs=mainCoefs#cbind(AP=rnorm(ncol(X)),NP=rnorm(ncol(X)))
   probs=Estep(newdat$Z,newdat$Posttest,X,coefs)
   probs1=probs+1
   iter=0
   maxDiff=numeric(200)
   conv <- 1
   while(!all(abs(probs1 - probs) < tol)){
      iter=iter+1
      probs1=probs
      invisible(capture.output(coefs<-Mstep(multDat,Xmult,probs)))
      probs=Estep(newdat$Z,newdat$Posttest,X,coefs)
      maxDiff[iter]=max(abs(probs1-probs))
      if(iter>200){
         conv <- maxDiff[iter]
         probs1=probs
      }
   }
   effDataNew=subset(newdat,Posttest==1)
   effDataNew$P=probs[newdat$Posttest==1, 'CP']
   effModNew=lm(Y~.+virtual:accelerated-Posttest-virtual-condition_assignment,data=effDataNew)
   eff <- unname(coef(effModNew)['Z'])

   ObsPostMod=lm(Y~.+virtual:accelerated-Posttest-virtual-condition_assignment-
P,data=effDataNew)

   obsPostEff=unname(coef(ObsPostMod)['Z'])

   return(c(eff=eff,obsPostEff=obsPostEff,conv=conv,iter=iter))
```

```
}

bsRes <- vapply(1:B,est1,est1(1))#,mc.cores=ncores)
save(list=ls(),file='results.RData')

effs <- bsRes['eff',]
obsPostEffs <- bsRes['obsPostEff',]
conv <- bsRes['conv',]

print(Sys.time()-begin)

save(list=ls(),file='results.RData')

save(mainCoefs,mainProbs,eff.est,obs.est,effs,obsPostEffs,file='bsResults.RData')
```