STAT 31900 — Assignment 1: Causal Inference Theories and Applications

Robert Winter

Table of Contents

1 Introduction **1**
- 1.1 Context . . . 2
- 1.2 Notation . . . 2
- 1.3 Research Questions . . . 3
- 1.4 Covariates . . . 3
- 1.5 Data Structure . . . 3

2 Analysis **3**
- 2.1 General . . . 3
  - 2.1.1 Question 1 . . . 3
- 2.2 Research Question 1 (Questions 2 – 4) . . . 5
  - 2.2.1 Question 2 . . . 5
  - 2.2.2 Question 3 . . . 6
  - 2.2.3 Question 4 . . . 7
- 2.3 Research Question 2 (Questions 5 – 10) . . . 10
  - 2.3.1 Question 5 . . . 10
  - 2.3.2 Question 6 . . . 12
  - 2.3.3 Question 7 . . . 13
  - 2.3.4 Question 8 . . . 15
  - 2.3.5 Question 9 . . . 17
  - 2.3.6 Question 10 . . . 19
- 2.4 Bonus Questions . . . 21

1 Introduction

Throughout this document, ***bolded*** *language signifies the problem set's instructions. My responses are provided in non-bold text.*

1.1 Context

All three assignments in this class are organized around investigations of the causal effect of class size reduction on student learning. The data for this first assignment came from a randomized experiment conducted in the State of Tennessee in 1985. Within each of the 76 participating schools, students entering kindergarten were assigned at random to one of three class types: a small class designed to have an enrollment range of 13–17 students; a regular class with an enrollment of 22–25 students; or a regular class with a teaching aide. Teachers were assigned at random to classes as well. Students were expected to remain in their initially assigned treatment conditions for four years. New teachers were assigned at random to classes in each subsequent grade. Students who did not attend kindergarten in the participating schools were randomized to different class types when they joined the study from the beginning of grade 1. Because past research has found no distinction in effectiveness between "regular class" and "regular class with a teaching aide," these are combined into one category.

More background information about this study and access to the entire data set can be found on this website: https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766.

Please find on Canvas a subset of the original Project STAR (Student-Teacher Achievement Ratio) data for this assignment. Please restrict the analysis to the 6,258 students who had valid information about treatment group membership in grade 1. This subsample includes the 2,313 students who joined the study in grade 1.

1.2 Notation

You may use $Z_k = 0, 1$ to denote the treatment assignment in kindergarten to a small class or a regular class, respectively, and use $Z_F = 0, 1$ for the corresponding treatment assignment in Grade 1; use Y_K to denote the kindergarten math score and Y_F for the Grade 1 math score.

Please generate a binary indicator R that takes value 1 if a kindergartner did not have the Grade-1 math score Y_F; let $R = 0$ if a kindergartner had a valid Grade-1 math score.

Note that for ease of interpretation, I reverse the treatment assignment variable, so that $Z_K = 0$ denotes assignment to a regular-sized kindergarten class and $Z_K = 1$ denotes assignment to a small kindergarten class. Similarly, $Z_F = 0$ denotes assignment to a regular-sized Grade 1 class and $Z_F = 1$ denotes assignment to a small Grade 1 class.

1.3 Research Questions

In this assignment, you are asked to evaluate two causal effects:

1. ***Research Question 1*: What is the causal effect of class size reduction in kindergarten (Z_K) on student *math* achievement by the end of Grade 1 (Y_F)?**
2. ***Research Question 2*: What is the causal effect of class size reduction in Grade 1 (Z_F) on student *math* achievement by the end of Grade 1 (Y_F)?**

1.4 Covariates

Observed baseline covariates include student gender, student race (you may combine Asians, Hispanics, American Indians, and the racial category labeled as "other" into a single category because the number of observations was relatively small for each of these racial/ethnic groups), and student free-lunch status (assuming that a family's financial situation did not differ between the kindergarten year and the first-grade year). Other covariates of potential interest include Grade 1 teacher career ladder level, and Grade 1 teacher's teaching experience.

In preparation for analysis, please see the Appendix [later in] this assignment for guidance on how to handle missing data in the covariates. (Please briefly explain the necessary steps you have taken in data preparation.)

1.5 Data Structure

Clearly, students were nested in schools and hence those who attended the same school should not be viewed as independent observations. If you have learned multilevel modeling or mixed-effects models in the past, please feel free to go ahead and specify such models (remember to report the robust standard errors). If you do not have such prior knowledge, you may employ OLS regression instead and obtain cluster robust standard errors in Stata or R with clusters indicated by school IDs.

2 Analysis

2.1 General

2.1.1 Question 1

1. We ask you to formally define the two causal effects of interest corresponding to the two research questions. Please address the following three

sub-questions:

1a. What is the target population in each case?

1b. What are the treatment conditions in each case?

1c. Define each population average causal effect of interest in terms of the expected values of the potential outcomes and explain every mathematical term in words in this application context.

Research Question 1

- The target population consists of all students who entered the study during kindergarten.
- There are two treatment conditions under this research question: (1) being in a regular-sized kindergarten classroom (with or without a teaching aide), denoted $Z_K = 0$, or (2) being in a small kindergarten classroom, denoted $Z_K = 1$.
- The average causal effect of being assigned to a small kindergarten classroom rather than a regular-sized kindergarten classroom on Grade 1 math achievement is $\delta^{(1)} = \mathbb{E}[Y_F(Z_K = 1)] - \mathbb{E}[Y_F(Z_K = 0)]$.[1] Here, $Y_F(Z_K = 1)$ is the distribution of the students' potential Grade 1 math scores in the counterfactual world where all students were assigned to small kindergarten classrooms; as such, $\mathbb{E}[Y_F(Z_K = 1)]$ is the population average potential Grade 1 math score associated with being in a small kindergarten classroom. Similarly, $Y_F(Z_K = 0)$ is the distribution of the students' potential Grade 1 math scores in the counterfactual world where all students were assigned to regular-sized kindergarten classrooms; as such, $\mathbb{E}[Y_F(Z_K = 0)]$ is the population average potential Grade 1 math score associated with being in a regular-sized kindergarten classroom. Thus, $\delta^{(1)}$ is the average effect of being in a small kindergarten classroom rather than a regular-sized kindergarten classroom on Grade 1 math scores, among all students who entered the study during kindergarten.

Research Question 2

- The target population consists of all students who entered the study during either kindergarten *or* Grade 1.
- There are two treatment conditions under this research question: (1) being in a regular-sized Grade 1 classroom (with or without a teaching aide), denoted $Z_F = 0$, or (2) being in a small Grade 1 classroom, denoted $Z_F = 1$.
- The average causal effect of being assigned to a small Grade 1 classroom rather than a regular-sized Grade 1 classroom on Grade 1 math achievement is $\delta^{(2)} = \mathbb{E}[Y_F(Z_F = 1)] - \mathbb{E}[Y_F(Z_F = 0)]$.[2] Here, $Y_F(Z_F = 1)$ is the distribution of the students' potential Grade 1 math scores in the counterfactual world where all students were assigned to

[1] We use the notation $\delta^{(1)}$ rather than simply δ to distinguish between the causal effects for Research Questions 1 and 2. The "(2)" should be read as corresponding to "Research Question 2," not as exponentiation.

[2] We use the notation $\delta^{(2)}$ rather than simply δ to distinguish between the causal effects for Research Questions 1 and 2. The "(1)" should be read as corresponding to "Research Question 1," not as exponentiation.

small Grade 1 classrooms; as such, $\mathbb{E}[Y_F(Z_F = 1)]$ is the population average potential Grade 1 math score associated with being in a small Grade 1 classroom. Similarly, $Y_F(Z_F = 0)$ is the distribution of the students' potential Grade 1 math scores in the counterfactual world where all students were assigned to regular-sized Grade 1 classrooms; as such, $\mathbb{E}[Y_F(Z_F = 0)]$ is the population average potential Grade 1 math score associated with being in a regular-sized Grade 1 classroom. Thus, $\delta^{(2)}$ is the average effect of being in a small Grade 1 classroom rather than a regular-sized Grade 1 classroom on Grade 1 math scores, among all students who entered the study during kindergarten or Grade 1.

2.2 Research Question 1 (Questions 2 – 4)

2.2.1 Question 2

2. We need to determine whether the population average causal effect of class size reduction in kindergarten on Grade 1 math achievement can be identified—that is, whether it can be equated with observable quantities in the target population. This question has three sub-questions:

2a. If you decide to simply compute the mean difference in the Grade 1 math outcome (Y_F) between the treated kindergartners ($Z_K = 1$) and the untreated kindergartners ($Z_K = 0$) in the target population, please write down the *prima facie* effect and explain it in words in this context.

The *prima facie* effect is

$$\delta_{PF}^{(1)} = \mathbb{E}[Y_F|Z_K = 1] - \mathbb{E}[Y_F|Z_K = 0].$$

Intuitively, this is the difference between the average Grade 1 math score among students who were assigned to a small kindergarten classroom and the average Grade 1 math score among students who were assigned to a regular-sized kindergarten classroom.

2b. Under what assumptions would this *prima facie* effect identify the causal effect of interest under Research Question 1?

This *prima facie* effect identifies the causal effect of interest if each student's kindergarten class size Z_K is independent of (1) their potential Grade 1 math score if assigned to a regular-sized kindergarten classroom, $Y_F(Z_K = 0)$, and (2) their potential Grade 1 math score if assigned to a small kindergarten classroom, $Y_F(Z_K = 1)$. In other words, each student's potential Grade 1 math scores—whether they be in a small or regular-sized classroom—must have no bearing on the kindergarten class size to which they are assigned, and similarly, the kindergarten class size to which they are assigned must have no bearing on their potential math scores in either class size. For example, it must *not* be the case that the students who perform best in small classrooms are systematically assigned to the small classroom treatment.

2c. Why are these assumptions plausible (or implausible) in the current study?

In this study, kindergartners were randomly assigned to small or regular-sized classrooms. Assuming the researchers implemented random assignment correctly, this randomness all but guarantees that classroom assignment is independent of potential Grade 1 math scores. After all, if students are randomly assigned to regular or small classrooms with no regard for how they would perform in each classroom type, then there is no way that classroom size and potential Grade 1 math scores could depend on one another. As such, the ignorability assumption is plausible in this study, and $\delta_{PF}^{(1)}$ can be seen as an appropriate identification of $\delta^{(1)}$.

2.2.2 Question 3

3. We ask you to, first of all, conduct a naive analysis of this data and generate an estimate of the *prima facie* effect. It involves the following steps:

3a. Compute the sample mean difference in the Grade 1 math outcome between kindergarten students attending a small vs. a regular class (remember the latter includes students attending a regular class with a teaching aide). Exclude students whose Grade 1 math score was missing (i.e., $R = 1$). Report the result as a point estimate of the population average treatment effect along with its standard error.

We perform a simple linear regression of Grade 1 math scores (`tmathss1`) on kindergarten class size (treated as a categorical variable, with categories for small class size, regular class size, and missing class size/students who were not in the study during kindergarten). For this analysis, we also cluster standard errors by the school at which students attended kindergarten, since this is the school where the intervention whose effect we are measuring was implemented.

Under this analysis, the population average treatment effect is 9.793. Intuitively, this means that, on average, assignment to a small kindergarten classroom increased students' Grade 1 math scores by roughly 9.793 points relative to their counterparts in regular-sized kindergarten classrooms. The standard error of this estimate is 2.273.

3b. Interpret the standard error.

The standard error of the coefficient estimate on $Z_K = 1$ captures the uncertainty in our estimate of this coefficient. In general, the higher the standard error of the estimate, the wider the range of values in which we are confident the true coefficient value lies. For example, assuming that Grade 1 math score conditional on kindergarten classroom size is normally distributed (a common assumption in regression analysis), our standard error of 2.273 means that we can be ~68% confident that the true effect of being in a small kindergarten classroom on Grade 1 math score is somewhere in the interval $(9.793 - 2.273, 9.793 + 2.273) = (7.52, 12.066)$.

3c. Report the result of hypothesis testing and the meaning of the p value in this context.

Our point estimate of the coefficient on $Z_K = 1$ has a p value of 6.34×10^{-5}. This means that the probability of observing a disparity in Grade 1 math scores between students in regular-sized and small kindergarten classrooms that is greater than the disparity we actually observed is extremely small: 6.34×10^{-5}, or roughly 0.00634%! This means that we can reject the null hypothesis that kindergarten class size has no effect on Grade 1 math scores in favor of the alternative hypothesis that kindergarten class size does have an effect on Grade 1 math scores, at even the $\alpha = 0.001$ significance level.

3d. Also report the estimated effect size (using the standard deviation of the control group students' math outcome as the scaling unit) and the 95% confidence interval for the effect size.

We estimate the effect size using Glass's delta statistic, calculated as the difference in mean outcomes between the treatment and control groups divided by the standard deviation of the control group's outcomes. Here, Glass's delta equals 0.233, which tells us that the mean Grade 1 math score among students in small kindergarten classrooms is 0.233 standard deviations higher than the mean score among students in regular-sized kindergarten classrooms. A 95% confidence interval for this effect size (calculated using the pivot method) is (0.162, 0.305). It is noteworthy that this confidence interval does not contain 0, indicating that at the $\alpha = 0.05$ significance level, the Glass's delta is significantly different from 0.

3e. What is your tentative conclusion about the effect of class size reduction in kindergarten on Grade 1 math achievement?

Thus far, we have found that there is a positive, statistically significant effect of small (as opposed to regular) kindergarten classroom size on Grade 1 math achievement, and that this effect amounts to approximately one quarter of a standard deviation of improvement relative to students in regular-sized kindergarten classrooms. As such, we can tentatively conclude that kindergarten class size reduction *does* improve Grade 1 math achievement, but perhaps only by a modest amount.

2.2.3 Question 4

4. By the initial experimental design, the 3,945 students who joined the study in kindergarten were expected to remain in their initially assigned treatment conditions in the subsequent years. However, there was a potential possibility that the randomized experiment might have been contaminated due to non-random attrition or non-compliance, which would require an investigation. This question has two sub-questions:

4a. *Non-random attrition:* Does the attrition rate differ between the kindergarten treated group ($Z_K = 1$) and the untreated group ($Z_K = 0$)? Would this result violate the identification assumption that you stated in Question 2b? Among those who had valid Grade 1 math scores, does the composition of kindergartners in terms of gender, race/ethnicity, and free-lunch status differ between the kindergarten treated group ($Z_K = 1$) and the untreated group

($Z_K = 0$)? Would this result violate the identification assumption that you stated in Question 2b? Please explain why or why not.

Of the 2,701 students in the untreated group (regular-sized kindergarten classrooms), 61 did not have valid Grade 1 math scores, for an attrition rate of $\frac{61}{2701} \times 100\% \approx 2.258\%$. Meanwhile, of the 1,244 students in the treated group (small kindergarten classrooms), 25 did not have valid Grade 1 math scores, for an attrition rate of $\frac{25}{1244} \times 100\% \approx 2.010\%$. We perform a chi-squared test to see if these two proportions are statistically significantly different from one another and recover a p-value of 0.704. This means that there is not statistical evidence that attrition rates were differential between the treated and untreated groups, which lends plausibility to the assumption of random attrition. Random attrition does not undermine our previously-stated ignorability assumption: the distributions of potential math scores within each treatment condition may change from their original distributions, but there is no indication that one treatment group's distributions would change in a systematically different way than the other's. (If, however, attrition *was* nonrandom, our assumption that treatment and Grade 1 math outcomes are independent would be violated. For instance, suppose attrition was differentially higher in the untreated group than in the treated group. In this case, students in regular-sized classrooms would be systematically exiting the study, perhaps because they were falling behind and their parents wanted them in a smaller classroom in a different school district. Then membership in regular-sized classrooms would be tied to student intelligence, which of course is tied to Grade 1 math scores. This clearly means that treatment and math score are not independent.)

Of the 2,640 students in the untreated group with valid math scores, 1,329 were female, for a female percentage of $\frac{1329}{2640} \times 100\% \approx 50.341\%$. Of the 1,219 students in the treated group with valid math scores, 604 were female, for a female percentage of $\frac{604}{1219} \times 100\% \approx 49.549\%$. We perform a chi-squared test to see if these two proportions are statistically significantly different from one another and recover a p-value of 0.6724. This means there is not statistical evidence that the treated and untreated groups had significantly different gender compositions after attrition, and both treatment groups after attrition were approximately 50/50 male/female.

Of the 2,640 students in the untreated group with valid math scores, 1,770 were white, for a white percentage of $\frac{1770}{2640} \times 100\% \approx 67.045\%$, and 861 were Black, for a Black percentage of $\frac{861}{2640} \times 100\% \approx 32.614\%$ (a *de minimis* percentage of students in the control group were of different racial/ethnic heritage). Of the 1,219 students in the treated group with valid math scores, 819 were white, for a white percentage of $\frac{819}{1219} \times 100\% \approx 67.186\%$, and 395 were Black, for a Black percentage of $\frac{395}{1219} \times 100\% \approx 32.404\%$ (a *de minimis* percentage of students in the treated group were of different racial/ethnic heritage). We perform a chi-squared test to see if these sets of proportions are statistically significantly different between the treated and untreated groups (due to the *de minimis* numbers of students who identified as something other than white or Black, our chi-squared test only compares the white and Black counts between the two treatment groups) and recover a p-value of 0.9373. This means there is not statistical evidence that the treated and untreated groups had significantly different racial/ethnic compositions after attrition, and both groups after attrition consisted of around $\frac{2}{3}$ white students and $\frac{1}{3}$ Black students.

Of the 2,640 students in the untreated group with valid math scores, at least 1,235 received free lunch assistance during Grade 1, for a free lunch percentage of around $\frac{1235}{2640} \times 100\% \approx 46.780\%$. Of the 1,219 students in the treated group with valid math scores, at least 591 received free lunch assistance during Grade 1, for a free lunch percentage of around $\frac{591}{1219} \times 100\% \approx 48.482\%$. (A small number of students in the treated and untreated groups had missing data regarding their free lunch status, so it is possible that the true free lunch percentages of each group are slightly higher.) We perform a chi-squared test to see if the proportions of free lunch recipients, non-free lunch recipients, and students with missing free lunch status are statistically significantly different between the treated and untreated groups and recover a p-value of 0.11. This means there is not statistical evidence that the treated and untreated groups had significantly different proportions of free lunch recipience after attrition, and both groups after attrition had just under 50% of students receiving free lunch assistance.

In summary, the demographic distributions (including gender/sex, race/ethnicity, and socioeconomic status, as measured by free lunch assistance) of students with valid Grade 1 math scores (i.e., after attrition) are not statistically significantly different between the treated and control groups.[3] This indicates that assignment of students to the two treatment groups was truly random with respect to demographic characteristics, and that attrition of students from each group was random with respect to demographic characteristics as well. This randomness suggests that there were no systematic differences between the students assigned to each treatment group, further supporting our ignorability assumption.

4b. *Non-compliance:* What proportion of students initially assigned to small classes in kindergarten ($Z_K = 1$) switched to regular classes in Grade 1 ($Z_F = 0$)? What proportion of those initially assigned to regular classes in kindergarten ($Z_K = 0$) switched to small classes in Grade 1 ($Z_F = 1$)? In each case, is noncompliance behavior predicted by kindergarten math achievement (Y_K)? (Please refer to the note on how to handle missing observations in a predictor.) Would this result violate the identification assumption that you stated in Question 2b? Please explain why or why not.

Of the 1,244 students assigned to small classes in kindergarten ($Z_K = 1$), 106 inappropriately enrolled in regular-sized classes in Grade 1 ($Z_F = 0$), for a treated group non-compliance rate of $\frac{106}{1244} \times 100\% \approx 8.521\%$. Of the 2,701 students assigned to regular-sized classes in kindergarten ($Z_K = 0$), 211 inappropriately enrolled in small classes in Grade 1 ($Z_F = 1$), for a control group non-compliance rate of $\frac{211}{2701} \times 100\% \approx 7.812\%$. We perform a chi-squared test to see if the proportions of noncompliant kindergartners are statistically significantly different between the treated and untreated groups and recover a p-value of 0.485. This means that there is not statistical evidence that the treated and untreated groups had significantly different rates of noncompliance, with both groups' noncompliance rates near 8%.

[3] This finding aligns with the conclusions of the authors of the original study, who wrote that "[t]he samples were compared on gender, race, and free-lunch composition to look for any systematic bias that may have arisen; none were found." See Finn, Boyd-Zaharias, Fish, and Gerber, *Project STAR and Beyond: Database User's Guide*, § 1.5.

To examine whether kindergarten math achievement predicts noncompliance among the treated group, we run a logistic regression of noncompliance on kindergarten math score. In particular, let $noncompliance_i = 1$ if $Z_K = Z_F = 1$ and $= 0$ if $Z_K = 1 \neq 0 = Z_F$. Also, to account for missing data, let $Y_{K\_miss,i}$ be an indicator for whether student i's kindergarten math score is missing, and let $Y_{K\_new,i} = Y_{K,i}$ if $Y_{K\_miss,i} = 0$ and equal the average kindergarten math score if $Y_{K\_miss,i} = 1$. We fit the logistic model

$$\mathbb{E}[noncompliance] = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 Y_{K\_new} + \beta_2 Y_{K\_miss}))}.$$

Using clustered standard errors (by kindergarten school, as we did in Exercise 3), we recover a p-value of 0.088 on our coefficient estimate of β_1 and a p-value of 0.999 on our coefficient estimate of β_2, which indicates that among students with non-missing kindergarten math scores, those math scores do not have a statistically significant effect on noncompliance within the treated group, and there is not a statistically significant difference in the likelihood of noncompliance between students with missing and non-missing scores. In short, then, there is not statistical evidence that kindergarten math score has an effect on noncompliance within the treated group.

To examine whether kindergarten math achievement predicts noncompliance among the control group, we run the same logistic regression model as above. Using clustered standard errors (by kindergarten school), we recover a p-value of 0.782 on β_1 and a p-value of 0.690 on β_2, which indicates that among students with non-missing kindergarten math scores, those math scores do not have a statistically significant effect on noncompliance within the untreated group, and there is not a statistically significance difference in the likelihood of noncompliance between students with missing and non-missing scores. In short, then, there is not statistical evidence that kindergarten math score has an effect on noncompliance within the untreated group.

In summary, we have found that in both the treated and and untreated groups, approximately 8% of kindergartners inappropriately enrolled in the "wrong" class size during Grade 1, meaning that noncompliance was (roughly) equally prevalent in both groups. We also found that in each treatment group, there was not a statistically significant systematic relationship between noncompliance and kindergarten math score—suggesting that when students *were* noncompliant and changed class sizes, this noncompliance had nothing to do with their potential Grade 1 math scores (which we would expect to be related to their actual kindergarten math scores). Taken together, this evidence supports our ignorability assumption, since it essentially suggests that noncompliance was random.

2.3 Research Question 2 (Questions 5 – 10)

2.3.1 Question 5

5. We now examine whether the population average causal effect of class size reduction in Grade 1 on Grade 1 math achievement can be identified. This

question [] has [four] sub-questions:

5a. If you simply compute the mean difference in the Grade 1 math outcome (Y_F) between the treated first-graders ($Z_F = 1$) and the untreated first-graders ($Z_F = 0$) in the target population, please write down the *prima facie* effect and explain it in words in this context.

The *prima facie* effect is

$$\delta^{(2)}_{PF} = \mathbb{E}[Y_F|Z_F = 1] - \mathbb{E}[Y_F|Z_F = 0].$$

Intuitively, this is the difference between the average Grade 1 math score among students who were assigned to a small Grade 1 classroom and the average Grade 1 math score among students who were assigned to a regular-sized Grade 1 classroom.

5b. Under what assumptions would this *prima facie* effect identify the causal effect of interest under Research Question [2]?

This *prima facie* effect identifies the causal effect of interest if each student's Grade 1 class size Z_F is independent of (1) their potential Grade 1 math score if assigned to a regular-sized Grade 1 classroom, $Y_F(Z_F = 0)$, and (2) their potential Grade 1 math score if assigned to a small Grade 1 classroom, $Y_F(Z_F = 1)$. In other words, each student's potential Grade 1 math scores—whether they be in a small or regular-sized classroom—must have no bearing on the Grade 1 class size to which they are assigned, and similarly, the Grade 1 class size to which they are assigned must have no bearing on their potential math scores in either class size.

5c. Why are these assumptions plausible (or implausible) in the current study?

In this study, students were randomly assigned to small or regular-sized Grade 1 classrooms before even beginning kindergarten. This randomization makes it highly plausible that Grade 1 classroom assignment was independent of students' potential Grade 1 math scores. After all, if students were randomly assigned to regular or small Grade 1 classrooms with no regard for how they would perform in each classroom type, then—as long as those students do not switch classroom types in Grade 1 in response to their performance in their assigned kindergarten classroom—there is no way that classroom size and potential Grade 1 math scores could depend on one another. As such, the ignorability assumption is plausible in this study, and $\delta^{(2)}_{PF}$ can be seen as a reasonable identification of $\delta^{(2)}$.

5d. Would your previous findings regarding nonrandom attrition in Question 4a or noncompliance in Question 4b arouse concerns about whether the class size effect in Grade 1 can still be identified by the *prima facie* effect? Why or why not?

In Question 4a, we found that attrition rates were comparable between students assigned to small and regular-sized kindergarten classrooms, and that the demographic distributions of the students who remained in the study through Grade 1 were comparable between the treatment groups. As such, there is no reason to believe that after random assignment of

students to the two treatment groups, attrition from or noncompliance with the study was disproportionately and systematically occuring in one group over the other. This means that, even taking attrition and noncompliance into account, our ignorability assumption is still plausible, and it is still reasonable to identify the Grade 1 class size effect with the *prima facie* effect described above.

2.3.2 Question 6

6. Conduct a naive analysis of the data and generate an estimate of the *prima facie* effect.

6a. Report the result as a point estimate of the population average treatment effect along with its standard error, the result of hypothesis testing, the estimated effect size and the 95% confidence interval for the effect size.

We perform a simple linear regression of Grade 1 math scores (`tmathss1`) on Grade 1 class size (treated as a categorical variable, with categories for small class size and regular class size). For this analysis, we also cluster standard errors by the school at which students attended Grade 1, since this is the school where the intervention whose effect we are measuring was implemented.

Under this analysis, the population average treatment effect is 11.446. Intuitively, this means that, on average, assignment to a small Grade 1 classroom increased students' Grade 1 math scores by roughly 11.446 points relative to their counterparts in regular-sized Grade 1 classrooms.

The standard error of this estimate is 2.277—a statistic that captures the uncertainty in our coefficient estimate. Assuming that Grade 1 math score conditional on Grade 1 classroom size is normally distributed, our standard error of 2.277 means that we can be ~68% confident that the true effect of being in a small Grade 1 classroom on Grade 1 math score is somewhere in the interval $(11.446 - 2.277, 11.446 + 2.277) = (9.169, 13.723)$.

Moreover, our point estimate has a p value of 4.4×10^{-6}. This means that the probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms that is greater than the disparity we actually observed is extremely small: 4.4×10^{-6}, or roughly 4.4×10^{-4}%! This means that we can reject the null hypothesis that Grade 1 class size has no effect on Grade 1 math scores in favor of the alternative hypothesis that Grade 1 class size does have an effect on Grade 1 math scores, at even the $\alpha = 0.001$ significance level.

Using Glass's delta statistic, the effect size of being in a small Grade 1 classroom on Grade 1 math outcome is 0.275. This tells us that the mean Grade 1 math score among students in small Grade 1 classrooms is 0.275 standard deviations higher than the mean score among students in regular-sized Grade 1 classrooms. A 95% confidence interval for this effect size (calculated using the pivot method) is (0.216, 0.335). It is noteworthy that this confidence interval does not contain 0, indicating that at the $\alpha = 0.05$ significance level, the Glass's delta is significantly different from 0.

6b. What is your tentative conclusion about the effect of class size reduction in Grade 1 on Grade 1 math achievement?

Thus far, we have found that there is a positive, statistically significant effect of small (as opposed to regular) Grade 1 classroom size on Grade 1 math achievement, and that this effect amounts to approximately one quarter of a standard deviation of improvement relative to students in regular-sized Grade 1 classrooms. As such, we can tentatively conclude that Grade 1 class size reduction *does* improve Grade 1 math achievement, but perhaps only by a modest amount.

2.3.3 Question 7

7. We will empirically examine whether the treatment effect estimate obtained in Question 3 might have been confounded by any of the observed covariates—student gender, race/ethnicity, and free-lunch status, Grade 1 teacher career ladder level, and Grade 1 teacher's teaching experience. A confounding variable would show an association with the treatment and an association with the outcome under one or both treatment conditions. This question has two subquestions:

7a. Examine if any of the above pretreatment covariates or their missing indicators may have confounded your previous estimate of the Grade 1 class size effect. Show your evidence (referring to analytic skills you have gained from an intermediate statistics course).

To examine whether gender is a confounder, we examine the relationships between gender and (1) Grade 1 class size, and (2) Grade 1 math score. We perform two chi-squared tests comparing gender and Z_F, one that includes students with missing Grade 1 math scores, and one that only includes students with non-missing Grade 1 math scores. These tests yielded p-values of 0.215 and 0.707, respectively, neither of which are below 0.05. As such, there is not statistical evidence that gender is associated with treatment. We also perform a regression of Grade 1 math scores on gender (with categories for male and female; no students with non-missing math scores had missing gender) among students with non-missing math scores, using standard errors clustered by Grade 1 school. We recover a p-value of 0.939 on the female category. This means that there is not statistical evidence that gender is associated with our outcome of interest. As such, there is not evidence that gender has confounded our estimate of the Grade 1 class size effect.

To examine whether race/ethnicity is a confounder, we examine the relationships between race/ethnicity and (1) Grade 1 class size, and (2) Grade 1 math score. We perform two chi-squared tests comparing race/ethnicity and Z_F, one that includes students with missing Grade 1 math scores, and one that only includes students with non-missing Grade 1 math scores. These tests yielded p-values of 0.226 and 0.510, respectively, neither of which are below 0.05. As such, there is not statistical evidence that race/ethnicity is associated with treatment. We also perform a regression of Grade 1 math scores on race/ethnicity (with

categories for white, Black, other, and missing race/ethnicity) among students with non-missing math scores, using standard errors clustered by Grade 1 school. We recover a p value of 3.518×10^{-9} on the Black category, 0.971 on the other category, and 0.150 on the missing category. This means that the Grade 1 math scores of Black students are statistically significantly different than their white classmates (though there are not significant differences between white students and students of non-white/Black race or missing race in the data). However, because race/ethnicity was not associated with treatment (including for Black students), there is not evidence that race/ethnicity has confounded our estimate of the Grade 1 class size effect.

To examine whether free lunch assistance is a confounder, we examine the relationships between free lunch recipience and (1) Grade 1 class size, and (2) Grade 1 math score. We perform two chi-squared tests comparing free lunch status and Z_F, one that includes students with missing Grade 1 math scores, and one that only includes students with non-missing Grade 1 math scores. These tests yielded p-values of 0.009 and 0.004, respectively, both of which are below 0.01. As such, there is statistical evidence that free lunch is associated with treatment. We also perform a regression of Grade 1 math scores on free lunch status (with categories for no free lunch assistance, free lunch assistance, and missing free lunch status) among students with non-missing math scores, using standard errors clustered by Grade 1 school. We recover a p-value of 5.589×10^{-16} on the free lunch category and a p-value of 3.289×10^{-2} on the missing free lunch status category. This means there there is also statistical evidence that free lunch status is associated with our outcome of interest — and even having missing free lunch status in the data is associated with the outcome of interest. As such, there is evidence that free lunch status *has* confounded our estimate of the Grade 1 class size effect.

To examine whether Grade 1 teacher career ladder level is a confounder, we examine the relationships between Grade 1 teacher career ladder level and (1) Grade 1 class size, and (2) Grade 1 math score. We note that Grade 1 teacher career ladder level has a number of categories, including some that are not career ladder levels at all: "chose no[t] to be on career ladder," "apprentice," "probation," "ladder level 1," "[ladder] level 2," "[ladder] level 3," and missing data. We perform two chi-squared tests comparing Grade 1 teacher career ladder level/cateogry and Z_F, one that includes students with missing Grade 1 math scores, and one that only includes students with non-missing Grade 1 math scores. These tests yielded p-values of 8.598×10^{-11} and 2.261×10^{-9}, respectively, both of which are well below 0.001. As such, there is statistical evidence that teacher career ladder level/category is associated with treatment, so that teachers at different levels in their careers (or who have failed to disclose career level information at all) are systematically in different-sized Grade 1 classrooms. We also perform a regression of Grade 1 math scores on Grade 1 teacher career ladder level/category (with categories for "chose no[t] to be on career ladder," "apprentice," "probation," "ladder level 1," "[ladder] level 2," "[ladder] level 3," and missing data) among students with non-missing math scores, using standard errors clustered by Grade 1 school. We recover p-values of 0.631 on the "apprentice" category, 0.594 on "probation," 0.364 on "level 1," 0.938 on "level 2," 0.107 on "level 3," and 0.592 on missing data, all of which are above 0.05. This means that there is not a statistically significant difference in Grade 1 math outcomes between students taught by teachers in each of the above categories and

students taught by teachers who did not disclose their career ladder level. As such, there is not evidence that Grade 1 teacher career ladder level has confounded our estimate of the Grade 1 class size effect.

To examine whether Grade 1 math teacher years of experience is a confounder, we examine the relationships between Grade 1 math teacher years of experience and (1) Grade 1 class size, and (2) Grade 1 math score. We perform two logistic regressions of Grade 1 class size on teacher years of experience (using the mean years of experience for teachers with missing data) plus a dummy variable indicating whether a teacher's years of experience was missing, with one regression including students with missing Grade 1 math scores and the other only including students with non-missing Grade 1 math scores. In the first regression, we recover p-values of 0.130 and 0.939 on the (possibly imputed) years of experience and missing indicator variables, respectively, and in the second regression, we recover p-values of 0.093 and 0.925 on these variables. Since none of these are below 0.05, there is not statistical evidence that Grade 1 math teacher years of experience is associated with treatment. We also perform a regression of Grade 1 math scores on the teacher years of experience (again using the mean years of experience for teachers with missing data) plus a dummy variable indicating whether a teacher's years of experience was missing among students with non-missing math scores, using standard errors clustered by Grade 1 school. We recover p-values of 0.860 on years of experience and 5.591×10^{-11} on the missing data indicator, the latter of which tells us that there are significant differences in Grade 1 math outcomes for students whose teachers did and did not have missing years of experience data. However, since years of experience was not associated with treatment, there is not statistical evidence that teacher years of experience confounded our estimate of the Grade 1 class size effect.

7b. Based on the empirical evidence, do you think that you might have underestimated or overestimated the Grade 1 class size effect? Why?

In the discussion above, we only singled out free lunch assistance as a potential confounder of our estimate of the Grade 1 class size effect. In particular, we found that in a regression of Grade 1 math scores on free lunch recipience, having free lunch assistance was associated with a roughly 26.930-point decrease in Grade 1 math scores. We also found that students with free lunch assistance were overrepresented in the untreated group, with at least $\frac{2387}{2004+2387+134} \times 100\% \approx 52.8\%$ of students in the untreated group receiving such assistance and only $\frac{865}{832+865+36} \times 100\% \approx 49.9\%$ of students in the treated group receiving such assistance. As such, the distribution scores in the untreated group was "pulled down" by its disproportionate composition of students receiving free lunch assistance, leading us to overestimate the Grade 1 class size effect.

2.3.4 Question 8

8. Make covariance adjustment for the identified confounder(s) along with their missing indicators. This question has three sub-questions:

8a. Write down your regression model. What are the model-based assumptions invoked in this analysis?

We now consider the following regression model:

$$Y_{F,i} = \beta_0 + \beta_1 \cdot Z_{F,i} + \beta_2 \cdot free\_lunch_i + \beta_3 \cdot \mathbb{I}_{missing\_free\_lunch_i} + \varepsilon_i,$$

where

- $Y_{F,i}$ is student i's Grade 1 math score,
- β_0 is the average Grade 1 math score among students in the untreated group who did not receive free lunch assistance,
- $Z_{F,i}$ is an indicator variable equal to 1 if student i was in a small classroom and 0 if student i was in a regular classroom,
- β_1 is the average change in Grade 1 math score associated with being in a small classroom rather than a regular-sized classroom,
- $free\_lunch_i$ is an indicator variable equal to 1 if student i received free lunch assistance and 0 otherwise,
- β_2 is the average change in Grade 1 math score associated with receiving free lunch assistance,
- $\mathbb{I}_{missing\_free\_lunch_i}$ is an indicator variable equal to 1 if student i did not have free lunch status data and 0 otherwise,
- β_3 is the average change in Grade 1 math score associated with having missing free lunch data, and
- ε_i is the random error term.

As usual in linear regression, an analysis using this regression model assumes that (1) classroom size and free lunch status (including having missing free lunch status in the data) are linearly associated with Grade 1 math scores, (2) the data form a random sample, (3) none of $Z_{F,i}$, $free\_lunch_i$, or $\mathbb{I}_{missing\_free\_lunch_i}$ are constant across all individuals i or can be written as a linear combination of one another, and (4) ε_i, conditional on the values of $Z_{F,i}$, $free\_lunch_i$, and $\mathbb{I}_{missing\_free\_lunch_i}$, has expected value 0 and constant variance σ^2. In addition, to interpret this model causally, we assume that $free\_lunch_i$ and $\mathbb{I}_{missing\_free\_lunch_i}$ have causal effects on both $Z_{F,i}$ and $Y_{F,i}$, and that their effects on $Y_{F,i}$ must be separated out from the effects of $Z_{F,i}$ on $Y_{F,i}$.

8b. Report the new result in examining the relationship between Grade 1 class size reduction and Grade 1 math achievement (point estimate, standard error, t statistic, p-value, effect size estimate, and 95% confidence interval for the effect size).

We perform the regression analysis described above, clustering standard errors by the school at which students attended Grade 1, since this is the school where the intervention whose effect we are measuring was implemented.

Now, we recover a population average treatment effect of 10.583, which is smaller than the previously-reported treatment effect of 11.446. Intuitively, this means that, on average, assignment to a small Grade 1 classroom increased students' Grade 1 math scores by roughly 10.583 points relative to their counterparts in regular-sized Grade 1 classrooms, holding all else equal.

The standard error of this estimate is 2.177, which is slightly smaller than the standard error in our previous model. This means we can be confident that the population average treatment effect falls somewhere in a tighter range around 10.583 than we previously thought around 11.446.

Moreover, our point estimate has a t statistic of 4.861 and a corresponding p-value of 8.157×10^{-6}, which is nearly half the size of our previous p-value of 4.4×10^{-6}. This means that the probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms, controlling for free lunch status, that is greater than the disparity we actually observed is extremely small: 8.157×10^{-6}. As such, we can reject the null hypothesis that Grade 1 class size has no effect on Grade 1 math scores in favor of the alternative hypothesis that Grade 1 class size does have an effect on Grade 1 math scores, at even the $\alpha = 0.001$ significance level.

Using Glass's delta statistic, the effect size of being in a small Grade 1 classroom on Grade 1 math outcome is 0.255. This tells us that the mean Grade 1 math score among students in small Grade 1 classrooms is 0.255 standard deviations higher than the mean score among students in regular-sized Grade 1 classrooms. This is a modestly smaller effect size than we previously found (0.275). A 95% confidence interval for this effect size (calculated using the Wald method) is approximately (0.152, 0.357).

8c. Would you change your conclusion about the causal effect of Grade 1 class size reduction in light of the new evidence obtained from 8b?

Based on our latest empirical analysis, I would still conclude that there is a positive, statistically significant effect of small (as opposed to regular) Grade 1 classroom size on Grade 1 math achievement — however, this effect is smaller than our previous analysis indicated. In particular, we now estimate being in a small Grade 1 classroom to increase students' Grade 1 math scores by 10.583 points, roughly 1 point lower than our previous estimate of 11.446 points. This still amounts to approximately one quarter of a standard deviation of improvement relative to students in regular-sized Grade 1 classrooms.

2.3.5 Question 9

9. Now restrict your analysis to the 2,313 students who joined the study in Grade 1 (i.e., `cltypek == 9` and `cltype1 != 9`) and report new results as you did in Questions 6a and 6b. Do these new results answer the initial question about the population average causal effect of class size reduction in Grade 1 on Grade 1 math achievement? Why or why not?

We perform a simple linear regression of Grade 1 math scores (`tmathss1`) on Grade 1 class size (treated as a categorical variable, with categories for small class size and regular class size). (Because the question asks us to mirror our workflow in Questions 6a and 6b, I have not included our previously-identified confounder, free lunch status, in this model.) We also cluster standard errors by the school at which students attended Grade 1, since this is the school where the intervention whose effect we are measuring was implemented.

Under this analysis, the population average treatment effect is 6.392. Intuitively, this means that, on average, students who entered Project STAR schools in Grade 1 and were assigned to small classrooms had Grade 1 math scores approximately 6.392 points higher than their counterparts who were assigned to regular-sized Grade 1 classrooms.

The standard error of this estimate is 2.853, which is slightly larger than the standard error we recovered on our point estimate in Question 6. This means that there is a wider range around 6.392 in which we could confidently expect the population average treatment effect to be than the range we expected around our coefficient estimate in Question 6.

Moreover, our point estimate has a p-value of 0.030. This means that the probability of observing a disparity in Grade 1 math scores between students in regular-sized and small Grade 1 classrooms—all of whom entered Project STAR schools in Grade 1—that is greater than the disparity we actually observed is small: 0.030, or 3%. This means that, at the $\alpha = 0.05$ level, we can reject the null hypothesis that Grade 1 class size has no effect on Grade 1 math scores among students entering Project STAR schools during Grade 1 in favor of the alternative hypothesis that among such students, Grade 1 class size does have an effect on Grade 1 math scores.

Using Glass's delta statistic, the effect size of being in a small Grade 1 classroom on Grade 1 math outcome among students entering Project STAR schools in Grade 1 is 0.158. This tells us that the mean Grade 1 math score among students new to Project STAR schools in small Grade 1 classrooms is 0.158 standard deviations higher than the mean score among students new to Project STAR schools in regular-sized Grade 1 classrooms. A 95% confidence interval for this effect size (calculated using the pivot method) is approximately (0.045, 0.271).

Altogether, this evidence suggests that, among students who entered Project STAR classrooms in Grade 1, being assigned to a small classroom rather than a regular-sized classroom caused a statistically significant, if somewhat modest, increase in Grade 1 math scores, with this effect amounting to approximately 0.158 standard deviations of improvement relative to students in regular-sized classrooms. However, these results do *not* answer our initial question about the population average causal effect of class size reduction in Grade 1 on Grade 1 math achievement. As stated in Question 1, the target population of Research Question 2 consists of all students who entered the study during kindergarten or Grade 1. This analysis, however, exclusively studies students who entered the study during Grade 1—a subset of students who may be systematically different than those who entered the study during kindergarten. Indeed, students who entered Project STAR schools during Grade 1 rather than kindergarten likely moved to the Project STAR school district in between Grade 1 and kindergarten. Moving from one town to another can be precipitated by parents quitting/losing a job and relocating to find a new one, the transfer of a child's custody from one

guardian to another, a change in parental income allowing (or requiring) relocation to a more expensive (or cheaper) neighborhood, belonging to a military family that frequently moves around, or another financial/familial factor that could influence student math achievement. New students at a school may also experience social challenges like bullying that impede their learning. While some of these factors may also affect students who entered Project STAR schools in kindergarten, these factors (and others like them) are likely *systematically* more prevalent among students who entered Project STAR schools in Grade 1. Restricting analysis to this subpopulation of students who are (most likely) not representative of the entire population of research subjects introduces selection bias, which in turn makes it no longer reasonable to identify the population treatment effect with our coefficient estimate.

2.3.6 Question 10

10. When using covariance adjustment as you did in Question 8 to answer Research Question 2, a tricky problem is that many students attending a small class in Grade 1 had also attended a small class in kindergarten and might have benefited from kindergarten class size reduction. You are asked to further consider the following questions:

10a. When making covariance adjustment for potential confounders in Question 8, if the analyst adds the kindergarten class type (Z_K) and its missing indicator to the covariance adjustment. Suppose that there are no unmeasured confounders. What would be the causal meaning of the regression coefficient for Z_F?

Adding kindergarten class type Z_K to our regression analysis (and excluding the free lunch confounder we previously identified for the sake of brevity), we now consider the model

$$Y_{F,i} = \beta_0 + \beta_1 \cdot Z_{F,i} + \beta_2 \cdot Z_{K,i} + \beta_3 \cdot Z_{K\_missing,i} + \varepsilon_i,$$

where:

- $Z_{F,i}$ equals 1 if student i was in a small classroom in Grade 1 and 0 if student i was in a regular-sized classroom in Grade 1,
- $Z_{K,i}$ equals 1 if student i was in a small classroom in Grade 1 and 0 otherwise, and
- $Z_{K\_missing,i}$ equals 1 if student i's Grade 1 status was missing (e.g., if student i was not in a Project STAR school in Grade 1) and 0 otherwise.

The expected Grade 1 math score for a student who was in regular-sized kindergarten *and* Grade 1 classrooms is $\mathbb{E}[Y_F|Z_F = 0, Z_K = 0, Z_{K\_missing} = 0] = \beta_0$. Moreover, the expected Grade 1 math score for a student who was in a regular-sized kindergarten classroom but a small Grade 1 classroom is $\mathbb{E}[Y_F|Z_F = 1, Z_K = 0, Z_{K\_missing} = 0] = \beta_0 + \beta_1$. And, the expected Grade 1 math score for a student who was in small kindergarten and Grade 1 classrooms is $\mathbb{E}[Y_F|Z_F = 1, Z_K = 1, Z_{K\_missing} = 0] = \beta_0+\beta_1+\beta_2$, while the expected Grade 1 math score for a student who was in a small kindergarten classroom but a regular Grade 1

classroom (i.e., a noncompliant kindergartner) is $\mathbb{E}[Y_F|Z_F = 0, Z_K = 1, Z_{K\_missing} = 0] = \beta_0 + \beta_2$. Thus, $\beta_1 = (\beta_0 + \beta_1) - \beta_0 = \mathbb{E}[Y_F|Z_F = 1, Z_K = 0, Z_{K\_missing} = 0] - \mathbb{E}[Y_F|Z_F = 0, Z_K = 0, Z_{K\_missing} = 0]$. Similarly, $\beta_1 = (\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_2) = \mathbb{E}[Y_F|Z_F = 1, Z_K = 1, Z_{K\_missing} = 0] - \mathbb{E}[Y_F|Z_F = 0, Z_K = 1, Z_{K\_missing} = 0]$. Both of these equations highlight that β_1 is the average difference in Grade 1 math scores between students in small and regular-sized Grade 1 classrooms *who were previously in the same type (i.e., small or regular-sized) of kindergarten classroom.* In causal terms, β_1 is the average effect of being in a small Grade 1 classroom rather than a regular-sized Grade 1 classroom on Grade 1 math achievement, *conditional on kindergarten classroom size.* For students who were in regular-sized kindergarten classrooms, β_1 may be thought of as the average effect on Grade 1 math achievement of enrolling in a small Grade 1 class rather than a regular-sized Grade 1 class. For students who were already in small kindergarten classrooms, β_1 may be thought of as the *incremental* average effect on Grade 1 math achievement of enrolling in a small Grade 1 class rather than a regular-sized Grade 1 class, above and beyond any benefit already derived from being in the small kindergarten classroom. (For such a student, the total causal effect of being in small kindergarten and Grade 1 classrooms on Grade 1 math achievement would be $\beta_1 + \beta_2$.)

10b. Some other analysts argue that kindergarten math achievement (Y_K) is a confounder of the causal effect of Grade 1 class size reduction on Grade 1 math achievement. They propose to add Y_K and its missing indicator to the covariance adjustment. Do you agree with this analytic plan? Why or why not?

On the one hand, I appreciate these analysts' concerns: kindergarten math achievement is certainly a predictor of Grade 1 math achievement (e.g., students who excel in math in kindergarten will likely excel in math in Grade 1). However, given our findings of random assignment, random attrition, and random noncompliance, we would expect "natural mathematical talent" (or any lack thereof) to be evenly distributed between the treated and untreated groups, so that kindergarten math achievement (Y_K) is not associated with Grade 1 class size (Z_F), in which case Y_K would not be a confounder of the causal effect of Z_F on Y_F. In turn, these analysts might reply that even though "natural mathematical talent" is evenly distributed between the treated and untreated groups, if small kindergarten classrooms really do contribute to higher (or lower) kindergarten math achievement, then since—assuming negligible noncompliance—kindergarten classroom size perfectly predicts Grade 1 classroom size, there *is* an association between Y_K and Z_F, meaning Y_K *is* a confounder that must be controlled for. Even in this case, however, there is a problem. Since—again assuming negligible noncompliance—Grade 1 classroom size is "locked in" for a kindergartner as soon as she enters a Project STAR school, there is a sense in which Y_K is a post-treatment variable: kindergarten math scores are an intermediate outcome between beginning enrollment in small classrooms and earning a Grade 1 math score. As such, I would not include kindergarten math outcomes as a covariate in my analysis.

2.4 Bonus Questions

B1. Derive the two sources of selection bias shown in Holland (1986) (see Week 1 Slide 28).

The bias from estimating the population average treatment effect δ by the *prima facie* effect δ_{PF} is given by:

$$
\begin{aligned}
\delta_{PF} - \delta &= (\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]) - (\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]) \\
&= (\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(0)|Z=0]) - (\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]) \\
(1) &= \mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(0)|Z=0] \\
&\quad - (\mathrm{P}(Z=0)\cdot\mathbb{E}[Y(1)|Z=0] + \mathrm{P}(Z=1)\cdot\mathbb{E}[Y(1)|Z=1]) \\
&\quad + (\mathrm{P}(Z=0)\cdot\mathbb{E}[Y(0)|Z=0] + \mathrm{P}(Z=1)\cdot\mathbb{E}[Y(0)|Z=1]) \\
&= (\mathbb{E}[Y(1)|Z=1] - \mathrm{P}(Z=1)\cdot\mathbb{E}[Y(1)|Z=1]) \\
&\quad + (-\mathbb{E}[Y(0)|Z=0] + \mathrm{P}(Z=0)\cdot\mathbb{E}[Y(0)|Z=0]) \\
&\quad - \mathrm{P}(Z=0)\cdot\mathbb{E}[Y(1)|Z=0] + \mathrm{P}(Z=1)\cdot\mathbb{E}[Y(0)|Z=1] \\
&= (1-\mathrm{P}(Z=1))\cdot\mathbb{E}[Y(1)|Z=1] - (1-\mathrm{P}(Z=0))\cdot\mathbb{E}[Y(0)|Z=0] \\
&\quad - \mathrm{P}(Z=0)\cdot\mathbb{E}[Y(1)|Z=0] + \mathrm{P}(Z=1)\cdot\mathbb{E}[Y(0)|Z=1] \\
&= \mathrm{P}(Z=0)\cdot\mathbb{E}[Y(1)|Z=1] - \mathrm{P}(Z=1)\cdot\mathbb{E}[Y(0)|Z=0] \\
&\quad - \mathrm{P}(Z=0)\cdot\mathbb{E}[Y(1)|Z=0] + \mathrm{P}(Z=1)\cdot\mathbb{E}[Y(0)|Z=1] \\
&= \mathrm{P}(Z=0)\cdot(\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(1)|Z=0]) \\
&\quad + \mathrm{P}(Z=1)\cdot(\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]) \\
(2) &= \mathrm{P}(Z=0)\cdot(\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(1)|Z=0]) \\
&\quad + \mathrm{P}(Z=1)\cdot(\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]) \\
&\quad + \mathrm{P}(Z=0)\cdot(\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]) \\
&\quad - \mathrm{P}(Z=0)\cdot(\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]) \\
&= (\mathrm{P}(Z=1) + \mathrm{P}(Z=0))\cdot(\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]) \\
&\quad + \mathrm{P}(Z=0)\cdot(\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(1)|Z=0] - \mathbb{E}[Y(0)|Z=1] + \mathbb{E}[Y(0)|Z=0]) \\
(3) &= (\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]) \\
&\quad + \mathrm{P}(Z=0)\cdot(\mathbb{E}[Y(1)-Y(0)|Z=1] - \mathbb{E}[Y(1)-Y(0)|Z=0]) \\
&= (\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]) \\
&\quad + \mathrm{P}(Z=0)\cdot(\mathbb{E}[\Delta|Z=1] - \mathbb{E}[\Delta|Z=0]),
\end{aligned}
$$

as desired, where (1) is by the Law of Total Expectation, (2) is by adding 0 in a "peculiar" way, and (3) is by linearity of expectation.

B2. Prove that each of these two sources of bias will become zero under independence.

If Z is independent of $Y(0)$ and $Y(1)$, then

$$(1) \quad \mathbb{E}[Y(0)] = \mathbb{E}[Y(0)|Z=0] = \mathbb{E}[Y(0)|Z=1], \text{ and}$$
$$(2) \quad \mathbb{E}[Y(1)] = \mathbb{E}[Y(1)|Z=0] = \mathbb{E}[Y(1)|Z=1].$$

Subtracting (1) from (2), we also have

$$(3) \quad \mathbb{E}[\Delta] = \mathbb{E}[\Delta|Z=0] = \mathbb{E}[\Delta|Z=1].$$

Using (1), our first source of bias (between-group differences in the potential outcome associated with the control condition) is

$$\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0] = \mathbb{E}[Y(0)] - \mathbb{E}[Y(0)] = 0.$$

Similarly, using (3), our second source of bias (between-group difference in the treatment effect) is

$$\mathrm{P}(Z=0) \cdot (\mathbb{E}[\Delta|Z=1] - \mathbb{E}[\Delta|Z=0]) = \mathrm{P}(Z=0) \cdot (\mathbb{E}[\Delta] - \mathbb{E}[\Delta]) = 0.$$

Thus, both sources of bias become zero under independence of Z and Y, as desired.

B3. Show that when using the *prima facie* causal effect to evaluate the average treatment effect on the treated (ATT), only the first source of bias possibly exists.

The bias from estimating the average treatment effect on the treated ATT by the *prima facie* effect δ_{PF} is given by:

$$\begin{aligned}
\delta_{PF} - \mathrm{ATT} &= (\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]) - \mathbb{E}[Y(1) - Y(0)|Z=1] \\
&= (\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(0)|Z=0]) - (\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(0)|Z=1]) \\
&= \mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0],
\end{aligned}$$

which is precisely the form of the first source of bias (between-group difference in the potential outcome associated with the control condition).