Sniffing Out the Secret Poison: Selection Bias in Educational Research

Daniel A. Showalter
Luke B. Mullet
Eastern Mennonite University

Selection bias is a persistent, and often hidden, problem in educational research. It is the primary obstacle standing in between increasingly available large education datasets and the ability to make valid causal inferences to inform policymaking, research, and practice (Stuart, 2010). This article provides an accessible discussion on the importance of understanding selection bias in educational research. Although a general explanation on how to remove selection bias is beyond the scope of this article, the reader is guided through an example of this removal process. Specifically, a propensity score analysis is used on a nationally representative dataset to examine whether high school course taking in the algebra-calculus pipeline has a causal effect on placing out of postsecondary remedial mathematics. Several visualizations of the selection bias, and the process of its removal, are provided to give readers a sense of its impact on analyzing observational data.

Introduction

Consider this fundamental problem in educational research: We would like to estimate the effectiveness of a particular policy, initiative, or pedagogical approach, but it would be impractical or unethical to conduct a true experiment by randomly assigning students to a group for the sake of research. Thus, we leave students in their existing groups, collect observational data on some variables of interest, and then run some correlational analyses to look for patterns. Finally, we report our results responsibly, reminding our readers that the results that we discovered should not be interpreted as implying causality, but rather a correlation.

Switch perspectives now to the eventual reader of the study who needs to make a decision about a particular policy, initiative, or pedagogical approach and is combing the research literature for advice. The reader comes across our study, finds the results attractive, and optimistically treats them as if we had indeed discovered a causal effect. One of the greatest problems with this process is *selection bias*.

Selection bias occurs when differences in average outcomes between treatment groups stem from differences in characteristics between group members rather than from the treatment. For example, say we want to determine the relationship between taking private hockey lessons and a hockey player’s odds of playing professionally. Basic statistics would show a correlation between these two variables—hockey players who take private lessons are more likely to play professionally than those who do not. However, there is strong potential for selection bias because the characteristics and opportunities of hockey players who take private lessons are substantially different than those who do not. For example, the average strength, speed, and coordination of those who start taking private hockey lessons is likely greater than the average strength, speed, and coordination of the general public. This stronger, faster, more coordinated group is more likely to become professional hockey players—regardless of whether or not they take private lessons. Therefore, we should not jump to the conclusion that there is a causal relationship before somehow isolating the effect of the lessons.

Selection bias has been described as the primary threat to validity when attempting to make causal claims with observational data (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007). On the other hand, unless educational research can make valid causal claims, the temptation is just to make decisions based on correlational results instead—a practice which can result in seriously flawed decisions (Stuart, 2010). For many years, educational researchers claimed that a causal question could be answered adequately only with a true experiment (e.g., see Light, Singer, & Willett, 1990). Fortunately, this assertion has been shifting as causal inference methods have been designed and refined for use with observational data. These newer methods attempt to assess causality when the research design does not include random assignment to treatment groups.

Reducing Selection Bias in Observational Data

Several statistical methods exist that can be used to help identify and reduce selection bias. For example, including covariates (variables that offer alternate explanations to why some groups have a better outcome than others) into a regression model has the potential to reduce selection bias, although it is often less reliable at identifying selection bias and less effective at removing it than other techniques (Rosenbaum & Rubin, 1983; Schneider et al., 2007). Other methods have been designed primarily to remove selection bias in order to make causal inferences with observational data. In a handbook published by the American Educational Research Association, Schneider et al. (2007) detailed four of these methods: fixed effects, instrumental variables, regression discontinuity, and propensity scores. A brief description of each method follows, but the interested reader is encouraged to consult Schneider et al.'s handbook for more thorough coverage.

Fixed Effects

Fixed effects methods attempt to account for influential variables that may be hard to observe or even identify but that do not change over time. This is particularly attractive when dealing with potential selection bias, because the model can effectively control for both observed and unobserved variables.

Consider the following scenario. A school principal is trying to decide between two after-school math programs, Mathatia and Numerate. Both programs have been used over the past several years, by the school's three math teachers. Pretests and posttests were administered each time one of the programs was offered. Mathatia seems to yield higher average improvement scores, but the difference is not statistically significant, and its curriculum is also more expensive. How can the principal get a better sense of how effective Mathatia actually is?

One solution would be to use a fixed effects model. It's likely that there is a substantial "teacher effect"; even using the same curriculum, some teachers will communicate the material more effectively than others. Rather than compare all the Mathatia courses with all the Numerate courses directly, we could separate the improvement scores by teacher. For example, what was the average improvement in Mathatia courses with Mrs. Mullet and how did this compare to the average improvement in Numerate courses with Mrs. Mullet? If we assume that Mrs. Mullet's teaching style is "fixed," this comparison would effectively control for all of the variables related to the teacher such as student-teacher relationships, teaching style, educational background, and content knowledge. Thus, some of the selection bias related to students selecting into Mrs. Mullet's class could be filtered out.

Fixed effects allows the researcher to control for both observed and unobserved variables that might cause selection bias. The disadvantages are that (a) it can be hard to find variables that are truly fixed (e.g., what if Mrs. Mullet gains experience over time and changes her teaching style?), (b) using fixed effects often drastically reduces the sample size, and (c) the results may not generalize to parts of the population.

Instrumental Variables

The instrumental variables approach attempts to identify a variable (the "instrument") that is highly correlated with the treatment but is uncorrelated with the outcome other than indirectly by way of the treatment (Leigh & Schembri, 2004). Statistical techniques can then be used to remove the variation in the treatment variable that is correlated with the instrumental variable. This allows the researcher to more precisely estimate how much of the treatment effect can reasonably be attributed to the treatment. The Venn diagram in Figure 1 shows the variation of an instrument, a treatment, and an outcome variable. A naïve statistical analysis might attribute everything in both the shaded and lined portions to an effect of the treatment on the outcome. However, part of this "effect" stems instead from selection bias, specification error, measurement error, and sampling error. The ideal instrumental variable (see the lined part of Figure 1) would filter out these error sources to yield the true effect of the treatment on the outcome.

This process is more easily understood with an example. Continuing with the example used to illustrate fixed effects, let's say that we wanted to estimate the effect of attending the Mathatia after-school program on improvement on a standardized math test. We need to find a variable that is highly correlated with whether a student joins the after-school program, but uncorrelated with how much their math score improves (except indirectly through their participation in the after-school program). This is not an easy task. Consider a variable such as whether or not the students have a parent or guardian who can pick them up after the program. This ride availability could be correlated with whether they enroll in the Mathatia program, but it would probably also be correlated with their math scores through variables such as home environment and family socioeconomic status. What if the Mathatia program offered free pizza as an incentive? Then, a student's preference for pizza might be a potential instrument. Students who enjoy pizza might be more likely to enroll in the program, and it is unlikely that preference for pizza would be correlated with math scores otherwise.

The primary advantage of using an instrumental variable model is that the removal of the excess variance reduces the selection bias. The main disadvantage is that it can be quite difficult to identify a variable that is highly correlated with the treatment and uncorrelated with the outcome variable (Schneider et al., 2007).

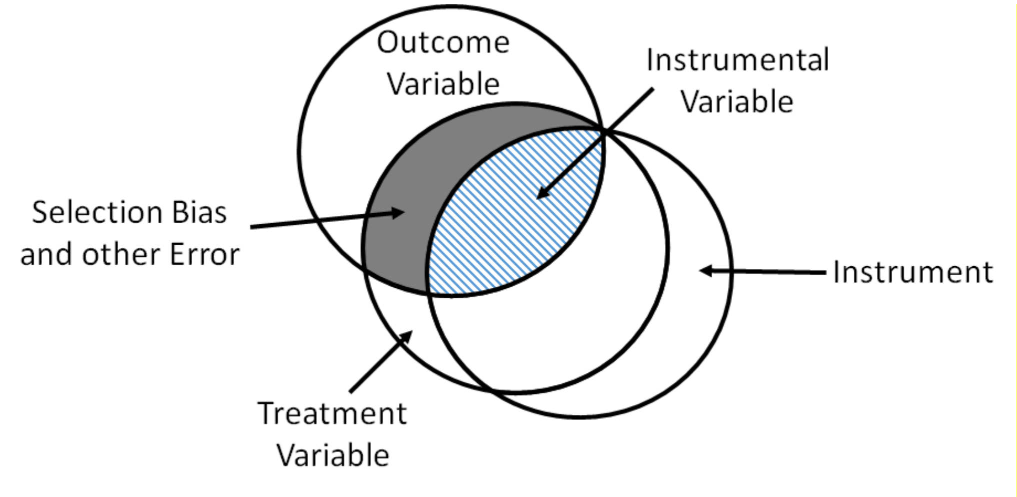

Figure 1. Venn diagram illustrating how an instrumental variable can filter out selection bias and other error sources.

Regression Discontinuity

A third method of causal inference, *regression discontinuity*, has been adopted increasingly in recent years. Regression discontinuity takes advantage of a situation where people below (or above) a certain cutoff point on a quantitative pretest are given a treatment.

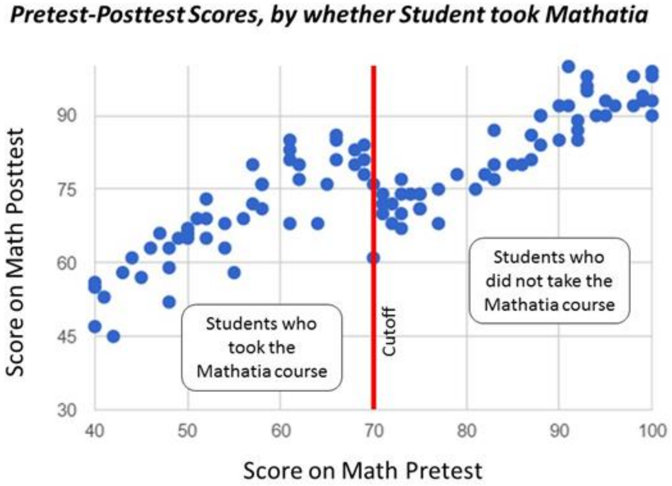

Figure 2. Fictional illustration of how regression discontinuity can be used to estimate a causal effect.

For example, say a group of students take a math pretest and are required to take a Mathatia course (the treatment) if their score was less than 70 (cutoff point). The students who scored above 70 do not take the Mathatia course. Then, the students take a math posttest. The regression slopes between the pretest and posttest performance for the two groups of students are then compared, and any “jump” that occurs at the cutoff point is attributed to the effect of the Mathatia course. Figure 2 demonstrates the discontinuity between the regression slopes of the two groups at the cutoff point. Similar to the fixed effects model, regression discontinuity attempts to cut through the potential variables causing selection bias and pinpoint the treatment effect.

Propensity Scores

Propensity score analysis combines a set of covariates into a single number that estimates how likely each participant was to receive the treatment. This estimated propensity score is then used to compare participants who received the treatment with other participants who had a similar propensity score but did not receive the treatment.

Returning to our after-school program example, propensity score analysis would mean analyzing numerous variables (such as gender, race, math ability, socioeconomic status, etc.) to predict how likely any given student was to enroll in the Mathatia program. Students who entered the program could then be compared with students who did *not* enter the program, but who had a similar likelihood of entering the program. In this way, students would be “fairly” compared to other students who were similar to them in gender, race, math ability, socioeconomic status, etc. If there were any differences in math improvement, we could attribute these effects to participating in the Mathatia program, since the students were similar on all the other variables we observed. Of course, some of these effects might be related to differences in variables that we didn’t observe, rather than the Mathatia program itself; this is a caution that one should always keep in mind when assessing how well a propensity score analysis did at removing selection bias.

Propensity score analysis typically works well with large datasets, where there are many observed variables for each individual. It is also helpful to have a large sample size that can afford to be reduced if it turns out that certain individuals are so different from others on the observed variables that they cannot reasonably be compared with each other.

Summary of Limitations on Using Causal Inference Methods

Just because causal inference methods such as the above are designed to reduce selection bias does not mean that they work well in every situation. The major disadvantage of using fixed effects methods is that there must exist multiple measurements of the outcome variable for each individual on which an effect is being estimated. This is impractical for many existing datasets and, when it is possible, can result in a substantial loss of the sample size (Aughinbaugh, 2012; Currie, 2003). Researchers who have attempted to use instrumental variables readily admit the delicate dependence of their results on the reliability of the instrumental variable (e.g., Altonji, 1995; Goodman, 2012; Rose & Betts, 2004). Regression discontinuity requires a standardized cutoff above which everyone is given a treatment and below which no one is given the treatment.[1] This scenario is relatively rare in practical settings. Propensity score analysis is impractical with small sample sizes and ineffective when there are not a sizable number of variables collected on each individual. Moreover, although propensity score analysis can

[1] There does exist a modified version of regression discontinuity known as *fuzzy regression discontinuity* which relaxes the assumption of a strict cutoff.

effectively remove selection bias from the observed variables, it only removes bias on the unobserved variables in as much as they are correlated with the observed variables.

In short, each of these four major causal inference methods has limitations on when it can be used. However, when the conditions are favorable, the methods offer powerful new options for filtering out selection bias and getting a more accurate picture of causation within educational research.

An Example with High School Mathematics Coursetaking

We now move to a timely example in educational research to truly understand how selection bias works, how a causal inference method can help remove it, and what the implications are of failing to remove it. Specifically, we will estimate the effect of high school students taking math courses in the algebra-calculus pipeline on the likelihood of placing out of remedial mathematics in college.

Background

Ever since algebra was first included in the public school curriculum in the 1800s, it has been a subject of debate (National Education Association, 1894; Reys & Reys, 2009). One of the most persistent debates centers on how many courses in the algebra-calculus pipeline all students should be required to take in high school. A major reason cited for requiring higher-level coursework is to reduce the billions of dollars spent on postsecondary remediation each year (Strong American Schools, 2008). Essentially, if students are required to take courses such as Algebra II and Precalculus in high school, they may be less likely to need remediation in postsecondary school—remediation that incurs financial costs (Scott-Clayton, Crosta, & Belfield, 2012; Strong American Schools, 2008 and psychological stigmas (Bailey, 2009), and allegedly decreases the chance of obtaining a degree (Adelman, 1999, 2006, Deil-Amen & Rosenbaum, 2002).

On the other hand, raising mathematics coursetaking requirements for high school graduation may have undesirable side effects such as lowering the quality of education for all students (Loveless, 2013) or raising high school dropout rates (St. John & Chung, 2006). Such risks might be tolerated more readily if mathematics coursetaking could be shown to have positive effects on postsecondary outcomes such as readiness for college mathematics.

Research has yet to demonstrate to what degree high school coursetaking in the algebra-calculus pipeline reduces the likelihood of postsecondary remediation, especially for students who would be unlikely to take upper-level courses in the pipeline unless they were required to do so in order to graduate. Moreover, whether or not to raise graduation requirements in mathematics has been a volatile issue in state policymaking recently; between 2004 and 2015, 19 states began requiring all students to take Algebra II (or even higher-level math courses), although two states (Texas and Florida) later retracted this requirement (Achieve, 2008; Kim, Kim, DesJardins, & McCall, 2015; J. Zinth, personal communication, April 25, 2016).

Previous Research

Studies have consistently shown that high school students who take upper-level courses in the algebra-calculus pipeline are less likely to need remedial mathematics in college (ACT, 2007; Adelman, 1999, 2006; Barber, 2011; Hoyt, 1999; Hoyt & Sorenson, 2001). The unanswered question is to what degree this reduced need for remediation comes from actually taking the mathematics coursework in high school and to what degree it is a byproduct of unrelated characteristics such as family socioeconomic status, mathematical aptitude, race, locale, or perseverance level. This is an important distinction, because if a sizeable effect is derived from the coursework itself, it would be much more logical to require all students to take the coursework as an act of social justice—especially if this effect was experienced even by students who were least likely to enroll in upper-level courses voluntarily. If, on the other hand, the difference is due to unrelated characteristics, then it would not make sense to force students to take the additional coursework (at least for the purpose of increasing their readiness for college mathematics).

One of the first studies to attempt to look beyond the obvious correlations to answer this question was done by Hoyt and Sorenson (2001). They used a logistic regression model to predict the likelihood of students in the Utah Valley State College system needing remedial mathematics coursework in college. As one of the independent variables in their model, the highest mathematics course taken in high school predicted 8% of the variance in a student's likelihood for needing mathematics remediation. Roth, Crans, Carter, Ariet, and Resnick (2001) also used a logistic regression model to study the question, but on a sample of students from Florida. They found that a student's highest-level mathematics course was a better predictor of reduced need for postsecondary mathematics remediation than was the student's overall high school GPA.

Consistent with these two localized studies, a third regression-based study on an even larger sample (n = 73,000) of Florida students found that the highest mathematics course taken in high school explained over 30% of the gap between whites and minorities, and between low-SES and high-SES students, in terms of need for postsecondary remediation (Long, Iatarola, & Conger, 2009). In addition to using a large sample, Long and his colleagues made some additional attempts to reduce selection bias through the use of fixed effects. Although these studies, especially that of Long et al., are suggestive of possible effects, they were each conducted within a single state. Single state research does not generalize well because states have a lot of power in determining educational policy. Due to policy differences, a possible relationship in a single state does not necessarily hold in another state. In addition, Long et al. (2009) made no clear attempt to show why the relationship was not just a correlation.

Barber (2011), using a nationally representative dataset, found that progression along the algebra-calculus pipeline was significantly associated with reduced need for postsecondary remedial mathematics even after controlling for a wide range of student- and school-level variables. However, aware of regression's inefficiency at adequately removing selection bias, she pointed out that her study should not be used as causal evidence of an effect.

Estimating a Causal Effect with a Large Dataset

To further inform state policymakers beyond what the previous research has done, we need to use a dataset that is both nationally representative and suitable for a statistical method that can remove selection bias. The National Center for Education Statistic's *Educational Longitudinal Study of 2002:06* (*ELS*) meets both of these conditions.

Dataset. The *ELS* comprises four waves of data collection. The first wave was in 2002, when a nationally representative sample of high school sophomores was selected. Each participant was interviewed and took standardized tests in reading and mathematics; a parent, English teacher, and mathematics teacher were interviewed for each student; and interviews were conducted with an administrator and a media specialist at each school. The follow-up waves involved longitudinal data collection on these same students in 2004, 2006, and 2012. A more complete description of the *ELS* can be found in U.S. Department of Education (2004).

Method I: (Mostly) ignore the selection bias. If we were to follow suit with much of educational research, we might do the following with the *ELS*. We could run a logistic regression (or perhaps some non-parametric equivalent) to see how well the highest math class a student took predicted whether they were placed into remedial math classes in college. Knowing that we want something more than a pure correlation, we would control for a few variables like socioeconomic status, standardized math ability, and perhaps a handful of others. Indeed, this would trim out a bit of the selection bias, but it would also entail a lot of guesswork and extrapolation as we will see later in the article. Nevertheless, we would have a result—probably a significant one based on what others have found (see, for example, Adelman, 2006 and Barber, 2011). So what does it really matter if it would contain some selection bias?

The problem would be that we would never have found a convincing answer to our question as to how effective it is to require students to take classes such as Algebra II or Precalculus in terms of their postsecondary preparedness. States might then assume that requiring all students to take Algebra II or Precalculus would improve postsecondary readiness among low-performing students and implement laws accordingly. Unfortunately, these decisions carry risks of increased dropout rates, higher barriers for struggling students, and a one-size-fits-all curriculum (Lillard & DeCicca, 2001; Loveless, 2013; St. John & Chung, 2006)—all without reasonable evidence of expected benefits for the students who were unlikely to enroll in these classes.

Method II: Account for selection bias. Of the four causal inference methods described earlier, propensity score analysis is a logical choice for answering our question since there are numerous variables collected on a sample of over 16,000 students. A general outline of propensity score analysis using a dataset such as the *ELS* is as follows:

1. Trim the dataset down to covariates related to the treatment and the outcome.
2. Use the remaining covariates to determine how likely each participant was to receive the treatment (i.e., their propensity score[2]).

[2] In the actual study, the logit of the estimated propensity score was used because of its statistical properties. However, the estimated propensity score is a much easier concept to grasp conceptually, and because the precise difference is irrelevant to the scope of this article, the estimated propensity score was used throughout the discussions.

3. Identify individuals with others who had a similar chance of receiving the treatment and then divide into groups based on which treatment they received.
4. Remove from the sample individuals who have no counterparts (individuals with similar propensity scores) in other treatment groups.
5. Run any desired statistical tests just as if the treatment groups had been assigned randomly.

In theory, if data could be collected on every possible variable, equivalent student groups could be created that would be even more balanced than what could be obtained through random assignment. In other words, we could use math to do more precisely what randomization does in the long run. Although the *ELS* certainly does not have data on every possible variable, it does provide substantial information from various perspectives on each student's home life, school activities, and personal characteristics. After removing students who are missing data either on the treatment variable (highest mathematics course taken in the algebra-calculus pipeline) or the outcome variable (whether the participant had taken any remedial mathematics coursework in college) we are left with an analytic sample of 9,416 students.[3] The goals of propensity score analysis are to take full advantage of the observed information on this sample to identify and group together students who are comparable on the observed variables. Returning to our hockey example, this would be like identifying a group of strong, fast, coordinated people who had taken private hockey lessons and a group of similarly strong, fast, coordinated people who had not. If the first group had become professional hockey players at a higher rate, then it probably was due to either the private lessons or to some unobserved variable (i.e., a variable other than speed, strength, or coordination).

Calculating the estimated propensity scores. The first task in reducing the selection bias is to select a subset of the variables to use as covariates in the propensity model. These are the "alternate explanations" for why students may place out of remedial math in college (other than due to their high school math coursework). In our case, we looked for covariates without too much missing data (less than 25%) that had theoretical connections to both the treatment and the outcome (Hirano, Imbens, & Ridder, 2003). For example, math self-efficacy has been tied to both coursetaking patterns and to placement in remedial courses in college, so we kept it as a potential covariate. This allowed us to boil the 2,604 variables in the public-use *ELS: 2002/06* dataset down to 230 variables. To select from among these 230 variables for our final model, we automatically included variables of strong theoretical interest, such as sex, race, socioeconomic status, and locale. Beyond that, we kept the variables with the 100 highest cross products—though somewhat arbitrary, this number is high enough to capture most of the variance related to the observed variables without creating a model that was overly complicated. It is worth noting that there was no need for each of these variables to contribute to the predictions in a statistically significant way; one of the advantages of propensity score analysis is that there is very little "cost" for each additional variable added to the model. It is much more important to find a set of covariates that works together to provide accurate estimates of the propensity scores.

[3] Most of the removed students were students who had never attended college, meaning that they were never even asked the question about taking remedial mathematics coursework in college.

The second step was to use our 108 covariates to estimate each student's propensity for eventually enrolling in Precalculus or higher during high school.[4] In a basic propensity score analysis, this would have entailed plugging our covariates into a logistic regression model. However, there were two hurdles to overcome in our case. The first hurdle was that we could have no missing data in order to run an ordinal logistic regression. Deleting every student who was missing data on even a single variable would have substantially hurt our ability to generalize our results, and we didn't want to do that! Thankfully, modern software offers several viable methods for dealing with missing data, such as multiple imputation and maximum likelihood. In our case, we used multiple imputation to fill in the missing data based on our information about each student (Allison, 2002; Graham, 2009).[5] The second hurdle was that our treatment was an ordinal variable (a categorical variable with an ordered scale of possible values), but the standard logistic regression model is based on a true/false outcome. In our case, there were several levels of mathematics a student could enroll in during high school: Prealgebra, Algebra I, Geometry, Algebra II, and Precalculus or higher. We handled this by running an ordinal logistic regression (a version of logistic regression that allows for an ordinal variable outcome). Specifically, we used SPSS to put our 108 variables in as predictors as well as the level of the highest math course taken as the outcome.

Our ordinal logistic regression provided us with an estimate of each student's propensity for taking Precalculus during high school.[6] Keep in mind that the dependent variable of the regression was the highest mathematics course taken during high school, and the independent variables were our 108 covariates. In other words, we had not touched anything to do with postsecondary remedial mathematics yet! Our regression was just a preliminary method to remove selection bias before we conducted the actual statistical tests.

Interpreting the estimated propensity score. Before proceeding, let's get a rough sense of what is represented by a propensity score. Consider a fictitious student, Pat. Knowing nothing about Pat—in essence, Pat could be absolutely any student who was a sophomore in a U.S. school in 2002—we would estimate Pat's propensity (tendency) to eventually take Precalculus as .580. We would guess this because the *ELS* tells us that 58% of the students in the country would eventually take Precalculus. Upon discovering that Pat is a female, we might adjust our estimate to .572 because the *ELS* shows us that females were slightly less likely to enroll in Precalculus than were males. Specifically, an estimated 57.2% of females who were sophomores in 2002 eventually took Precalculus in high school. However, we would raise our propensity score estimate to .614 when we find out that Pat attended high school in an urban area, because urban students were more likely to take Precalculus than non-urban students.

[4] The highest-level mathematics group consists of the students who enrolled for at least one semester in Trigonometry, Precalculus, or Calculus. For simplicity's sake, these students are referred to as the "Precalculus" students throughout this article.

[5] Any process used to fill in missing data should be used cautiously. In our case, we created 25 predictions for each missing data value, and then used the average of these predictions. We then ran a series of sensitivity analyses to evaluate the accuracy of these predictions and the potential impact on the estimated propensity scores.

[6] Although the data can now be analyzed by any method suitable for experimental data, the reader is reminded that the ability to make a causal claim depends on the assumption that the unobserved variables do not contain substantial selection bias.

The more information we learn about a student, the more accurately we are able to estimate their propensity for eventually taking Precalculus. As our estimate approaches the student's true propensity score, additional pieces of information tend to alter our estimate less and less. We become more confident that any selection bias contained in the unobserved variables would not substantially impact our estimates of a student's tendency to take Precalculus (Hong, 2004; Rubin, 1976). The value of the true propensity scores cannot be overstated; if we had access to the true propensity scores, we would be able to achieve balance in treatment groups on *all* variables just as well as random assignment does (Rosenbaum & Rubin, 1983).

What does it mean to have "balanced treatment groups"? Let us assume that, after accounting for Pat's data on the 108 covariates in our model, our estimate of her propensity score shifts from .614 to .602. If we were to scan our analytic sample of 9,416 students, it is quite possible we could find another student with a propensity score of .602. Would it be reasonable to infer that this new student was probably also a female who attended an urban school and shared many of the same characteristics as Pat? Not at all! However, say that we found a collection of 200 students with a propensity score of .602. Perhaps 60 of these students took courses through Precalculus, 40 took through Algebra II, 50 through Geometry, 30 through Algebra I, and 20 through Prealgebra. What we could say with reasonable confidence is that the proportion of females in the Precalculus group would be unlikely to differ significantly from the proportion of females in the Geometry group. And, the average standardized English test score of the Prealgebra group would be unlikely to differ significantly from that of the Algebra II group.

In other words, on average, groups of students who share the same propensity score are reasonably similar to each other on all of the covariates we included in our model, regardless of how far they eventually progressed in the algebra-calculus pipeline. How could this be? The details are beyond the scope of this article, but Rosenbaum and Rubin (1983) proved mathematically that this process achieves the same balance on the observed variables that would be possible if we could randomly assign each participant to a treatment group. The gist of their proof is that, in order to have the same propensity scores, students would have to have similar values on predictors that carried the most weight in the logistic regression. But *why* is it important to have balanced treatment groups? Let's say that two groups have essentially the same average value on every explanatory variable except that everyone in one group got Treatment A and everyone in the other group got Treatment B (or no treatment at all). If we then observe significant differences in the outcome between the two groups, the most logical explanation for the cause of the difference is the treatment itself. In other words, we have successfully extracted the selection bias!

In Pat's hypothetical scenario where we grouped together students with estimated propensity scores of .602, selection bias would be essentially removed and balance would be achieved between the groups in the algebra-calculus pipeline, at least in terms of the 108 covariates in our model. Of course, the magic of propensity scores only covers the observed variables; it's possible that the groups would be imbalanced on some variable that we didn't measure. This is the main justification for including a wide range of covariates with theoretical connections to the treatment and outcome in our model.

Removing selection bias to create balanced treatment groups. Unfortunately, Pat's example is more illustrative than realistic. There are two problems with attempting to match students with all other students sharing the same estimated propensity score. First, it is very possible that all (or none) of the students at a certain propensity score went beyond certain points in the algebra-calculus pipeline. Among our .602'ers, there may be several who took Precalculus, but none whose highest math course was Prealgebra. In fact, maybe none of the students who stopped at Prealgebra had high propensities to eventually reach Precalculus (we'll soon see that this was exactly what happened in our dataset). Second, it is highly unlikely that 200 of our 9,416 students have an estimated propensity score of .602, regardless of what course they ended up in. There's nothing magical about the hypothetical .602 either. In a large group of people, you may have two or three that were born on the day at the same time, but you probably won't have that many; in the same way, it is unlikely for large clusters of our students to have exactly the same propensity score.

The first problem is a serious one related to selection bias that is often overlooked in research. There are certain limits to what can be done with even the most powerful statistics. One of these limits is that we should not compare students who are too different from each other on important variables—at least when we're trying to estimate the effect of some treatment or program (Ho, Imai, King, & Stuart, 2007). If we do, we'll never know how much of our presumed "effect" is due to the treatment, and how much is due to the pre-existing differences between the students. What should be done instead? Since the goal is to minimize selection bias, the most recommended option is to remove the incomparable students and be sure to note the implications when generalizing the results to the population (Lechner, 2002).

The second problem of not finding enough students with a given propensity score is more manageable. One possibility is to group together all students with estimated propensity scores within a certain range, say from .600 to .650; the wider the range, the greater the sample size within each treatment level. The tradeoff is that, as our range widens, the differences between propensity scores increase, and our balance on the observed covariates steadily decreases. Statistical techniques exist to restore some of the balance such as inverse probability of treatment weighted (IPTW) and marginal mean weighting through stratification (MMW-S). Our use of MMW-S followed the steps described by Hong (2012).

The art of this process is in creating enough propensity score strata to achieve balance (and thus remove selection bias), but not so many that we lose statistical power by small or nonexistent sample sizes for each treatment level within any stratum. Figure 3 demonstrates hypothetically how increasing the number of strata increases balance on socioeconomic status between the Precalculus and Algebra 1 groups, while simultaneously decreasing the sample size within each stratum. The data here is a fictional set of 32 students. Broken down by SES, there are 10 in the upper third, 11 in the middle third, and 11 in the lower third. By highest math class taken in high school, 17 of these students took through Precalculus and 15 took through Algebra I. Each student's horizontal placement in each of the three strata scenarios is determined by their estimated propensity to eventually take Precalculus as of their sophomore year, when they were surveyed.

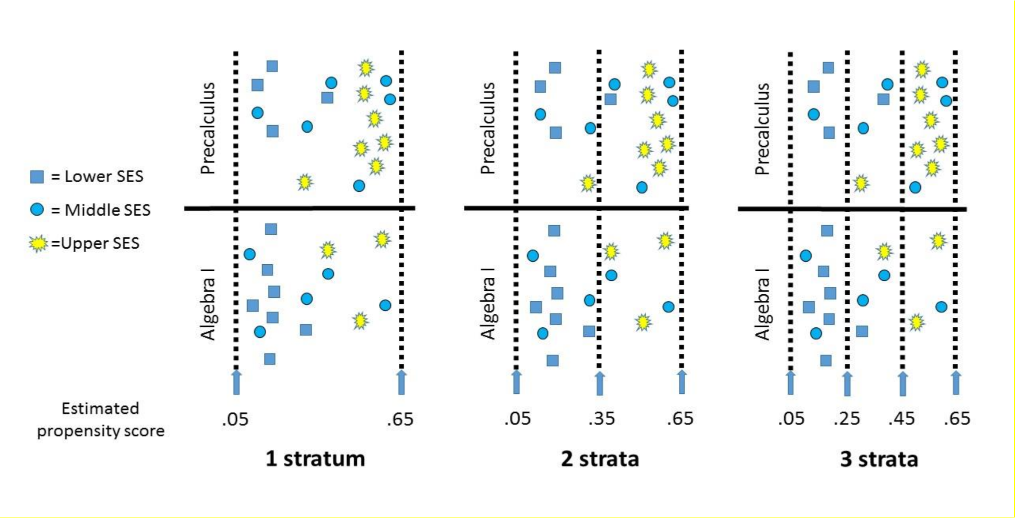

Figure 3. Achieving balance of socioeconomic representation within each stratum, by increasing the number of strata.

With just one strata, there is heavy imbalance between the top and the bottom in terms of proportion of students in each socioeconomic class. For example, about 41% of the students who completed through Precalculus were in the highest SES third, but only 20% of the students completing through Algebra were in the highest SES third. That difference opens the door to selection bias. With two strata, the SES balance gets better: In the lower of the 2 strata, 17% of the students who eventually took Precalculus are high SES (with 0% of Algebra as high SES), but in the upper of the 2 strata, the percentages are 55% and 60%. Finally, with three strata, the percentage of the eventual Precalculus students who are high SES in each strata are 0%, 25%, and 67%—the same as the respective percentages among the students who only completed through Algebra.

So, more strata equals more balance, with the caveat being that with too many strata, sample sizes are too small to run a statistical test for each stratum. With the selection bias removed, or at least greatly reduced, it is possible to analyze the groups for causal effects, just as if the treatment groups had been formed through random assignment.

Visualizing Selection Bias with Figures

Because selection bias is an abstract concept, it can be helpful to visualize how it takes shape in our data. Using the aforementioned process on our analytic sample from the *ELS*, we estimated a propensity score for each of our 9,416 students; the longitudinal nature of the *ELS* allows us to track these students and see what their actual highest mathematics course was in high school.

Selection bias in histograms. Figure 4 shows us how far students in the *ELS* were predicted to go in the algebra-calculus pipeline based on their covariate data, as well as how far they actually went. The top histogram shows the distribution of the 245 students whose highest mathematics course, by the end of high school, was Prealgebra, General Math, or Consumer Math. The further a bar is to the left, the less likely it was that the students represented by that bar would eventually take an upper-level mathematics course. For example, since the bars in the top histogram are positioned mostly to the left, we know that most of these students were unlikely to ever enroll in Precalculus. In contrast, the bottom histogram shows the distribution of the 5,038 students whose highest course was Trigonometry, Precalculus, or Calculus.

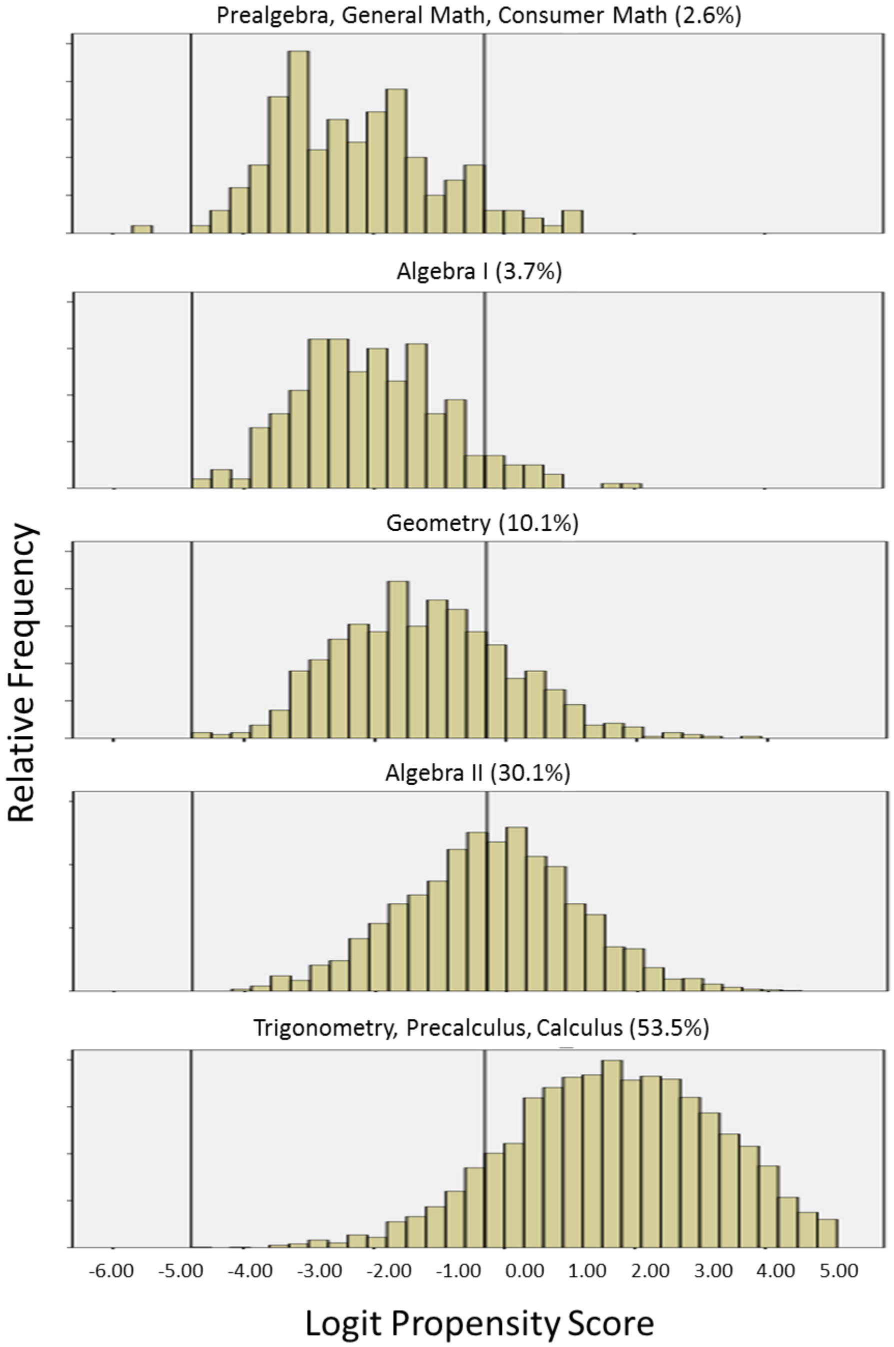

Figure 4. Distributions of logit propensity score by highest mathematics course taken. Histogram bars between the two vertical lines indicate relative frequency of students who can be compared to students in any of the other four groups based on similar pretreatment characteristics. The percentage of the analytic sample within each group is shown in parenthesis.

Three points are worth noting about Figure 4. First, the students who only completed through a lower-level math course had propensity score estimates that were quite different from those who took through an upper-level math course. In other words, for a majority of the students, knowing characteristics about these students early in high school would allow one to accurately predict how far they would progress in the math pipeline (or at least which category they probably would not end up in). Our ability to do this should raise some red flags that selection bias should be a concern.

Second, the greater the differences in propensity scores (i.e., the further apart the bars on the histograms), the higher the risk of selection bias distorting any actual effects; in particular, only the students in-between the vertical bars were deemed “comparable” by the research standards described earlier. Unfortunately, 6,245 students, or 66.3% of the analytic sample, were outside these boundaries. *Regardless of the statistical method used*, these students cannot be accurately analyzed for any potential treatment effect—but a researcher unaware of selection bias would have blindly included them in the study along with the remainder of the students who are similar enough to be compared. So, with regards to these 6,245 students, we’ve advanced from being unaware that we don’t know to *knowing* that we don’t know. While following in the footsteps of Socrates may bring some comfort, our situation is even more hopeful. We are also saying that we can accurately estimate an effect for the remaining 3,000+ students. In our case, this works out perfectly, because it’s precisely this portion of the population (i.e., students unlikely to take Precalculus) who we want to find out about anyway. They are the students who would be impacted by a policy change requiring them to take more upper-level math.

Third, a researcher using OLS regression or logistic regression, the methods that have often been used to conduct studies in this domain, would not have been able to see the data in this way. More importantly, they would erroneously have to make the assumptions that the shapes of these five distributions are essentially the same (Stuart, 2010). This would mean that it should be possible to draw a vertical line through Figure 4 that would come near the highest point of each of the five histograms. This is clearly not the case, indicating that selection bias should be a serious concern for any causal analysis of this entire group of students.

Selection bias in boxplots. The entire analytic sample of 9,416 is represented in Figure 4. The boxplots in Figure 5 show the distribution of propensity scores, within each highest math course category, only for the 3,171 statistically comparable students. These are the students who, based on characteristics present early in high school, could have plausibly ended up in any of the five categories along the algebra-calculus pipeline. However, even within this subsample, the groups of students were imbalanced. An ANOVA to test for differences in average propensity score between the treatment groups was significant, $F(4, 3166) = 178.869$, $p < .001$. If the five groups had been comparable, we would expect these five boxplots to look the same (i.e., the boxes would be of similar height and location on the vertical axis, and the whiskers would start and end at roughly the same vertical location). Clearly, this is not the case. The boxes are much higher in the higher mathematics groups (on the far right of Figure 5), meaning that, in their sophomore year, the students who eventually took an upper-level class such as Algebra II or Precalculus were already very different in categories such as race, gender, and teacher expectations than their peers whose eventual highest mathematics class was Prealgebra or Algebra I. To more precisely gauge these differences, we ran an ANOVA to test for differences on each of the covariates. Of the 108 ANOVA tests, only 3 were not statistically significant at an alpha level of .05. The other 105 contain a significant amount of imbalance between the students in different treatment groups. Therefore, even after identifying the students who could reasonably be compared to each other, we still had a major selection bias problem to address.

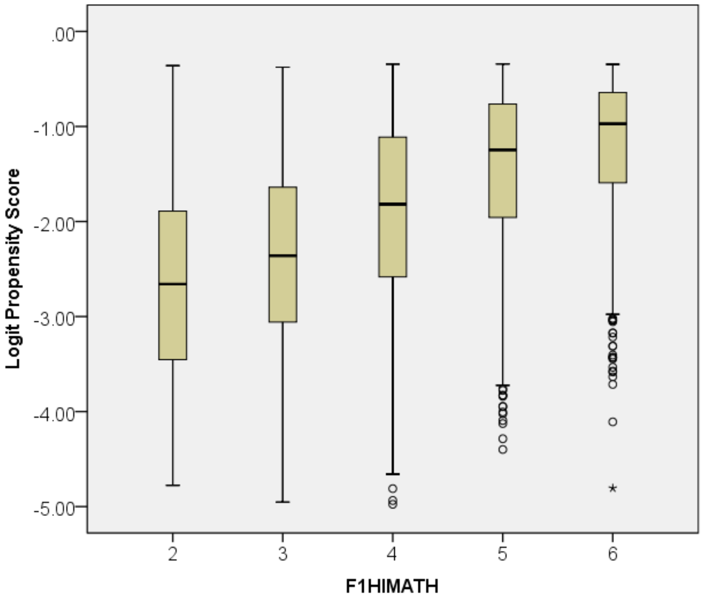

Figure 5. Distribution of logit propensity score by highest mathematics course taken in the analytic sample. If the five groups were balanced on all pretreatment characteristics (i.e., no selection bias), the boxplots would appear identical to each other.

However, there is an important difference between the imbalance shown among the entire analytic sample (in Figure 4) and the imbalance shown among only the comparable students (in Figure 5). The latter imbalance can be “fixed” using appropriate statistical methods, whereas the former imbalance is due to differences inherent in the students themselves and cannot be convincingly corrected for *regardless of the statistical method used*.

Because the 3,171 students in the subsample at least had a reasonable chance of ending up anywhere along the algebra-calculus pipeline, it is now possible to separate them into relatively balanced groups based on their estimated propensity scores. We can accomplish this balance by using increasingly smaller strata until there is no significant difference found between the estimated propensity scores among the five treatment groups within each strata. In this particular example, it took 11 strata to achieve this balance. At this point, marginal mean weights (Hong, 2010, 2012) were applied to privilege the valuable information provided by students such as low-propensity students who, against all odds, completed through Precalculus (or high-propensity students who took no classes above Prealgebra). The resulting five distributions are shown in Figure 6. The balance visualized here represents the overall balance when students are only compared to other students within their same propensity score stratum. A one-way ANOVA found no significant differences in average propensity score between treatment groups, *F*(4, 3165) = 0.066, *p* = .992.

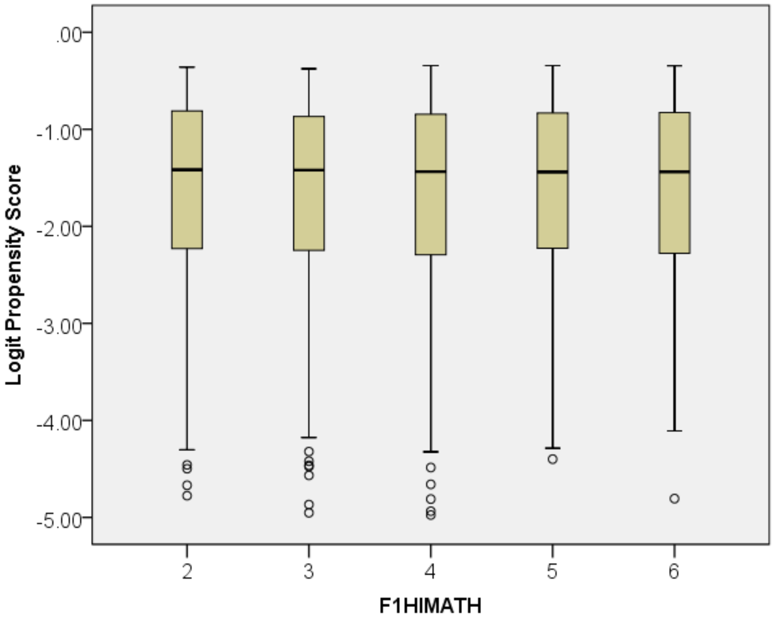

Figure 6. Distribution of logit propensity score by highest mathematics course taken after applying marginal mean weights. In contrast with the boxplots in Figure 5, the boxplots shown here are similar to each other, suggesting that selection bias has been essentially removed by applying the marginal mean weights.

Checking whether selection bias has been removed. An overall balance on the propensity scores does not guarantee balance between individual groups on any given variable. For example, within a particular stratum, is there a significant difference between the Prealgebra students and the Precalculus students in terms of the proportion that reported thinking math is fun, or the proportion of females, or how far their parents expect them to go in their education? Let's check!

In total, this resulted in over 10,000 pairwise comparisons, 92.1% of which had no significant difference between the group means. The results of the pairwise difference, by stratum, are presented in Table A1 of the Appendix. Hypothetically, if random assignment could have been used to assign students how far to go in the algebra-calculus pipeline, we would expect about 95% of these pairwise comparisons to be nonsignificant at an alpha level of .05. In general, our groupings did almost as good as random assignment, except for in the top two strata. Let's take a look at the results both with and without the effect sizes of these strata included.

Double checking whether selection bias has been removed. There are several ways to double check whether we adequately removed selection bias on the observed variables. Ho et al. (2007) recommended that standardized differences between means (e.g., Cohen's *d*) be calculated for the most influential predictors in the propensity model. In plain English, we want to see how different our group averages are on the important variables, and we want to see all of these differences on the same scale. To be safe, we should look at more than the averages; we should also make sure the data is roughly spread out in the same way for each variable (Imai, King, & Stuart, 2008). The standardized differences in means and variances can be found in Table A2 in the Appendix. If MMW-S has functioned as intended, all of the treatment groups within each propensity score stratum will have similar means and variances on influential predictors. There is no commonly-accepted standard on what constitutes "similar," but a Cohen's *d* of 0.25 or less is the standard used by the What Works Clearinghouse (2013) to indicate that groups may be reasonably compared.

Selection bias in the unobserved variables. We can now be confident that essentially all of the selection bias related to the observed covariates has been removed. But how much selection bias remains due to group differences in the unobserved variables? Unfortunately, we cannot be sure. This uncertainty is what separates propensity score analysis from a true experiment. However, Shadish, Clark, and Steiner (2008) devised an ingenuous method to test the validity of using causal inference methods, and propensity score analysis in particular, to make causal inferences. They randomly assigned a sample of 445 university students to a true experiment or to an observational study. They randomly assigned students in the true experiment to a training session in mathematics or to one in vocabulary, but they allowed the second group to select a training session. They then analyzed the experimental data and compared it with a variety of propensity score techniques on the observational data. The degree to which propensity score techniques removed selection bias varied wildly, with one type reducing only about half of the bias and another reducing about 95% of the bias; however, it is important to note that the former was a result of only including a few covariates in the propensity model. Shadish and his colleagues (2008) concluded that most of the propensity score techniques used in their study had been suitable for making causal inferences in their study (i.e., there wasn't much selection bias left after accounting for the bias in the observed variables), but recommended that researchers who use propensity score techniques include more than a small set of demographic variables. Lucky for us, we used our 108 covariates!

Analyzing the sample after selection bias has been removed. After removing the selection bias, it is possible to run any statistical test for an effect just as if the balance had been achieved by random assignment in an experiment.[6] Therefore, we can finally get to the actual statistical tests in our example. Essentially, this entails conducting a separate experiment within each stratum, and obtaining a separate effect size for each stratum. Because we have an ordinal-level treatment and a dichotomous outcome, we can calculate an effect size known as Kendall's tau b. A minimum practically significant effect would be roughly around a tau of 0.2 and a moderate effect would be about 0.5. Effect sizes for each stratum are presented in Table 1.

The average effect size for all 11 strata is tau = .04, 95% CI [–.10, .17]. Among the lower nine strata, for which we can be confident that the selection bias on the observed covariates has been removed, the average effect size is tau = .02, 95% CI [–.12, .16]. Among the lowest four strata, the average effect size is tau = –.03, 95% CI [–.19, .13]. Sensitivity analyses can help us gauge how much these effect sizes might differ if a portion of the students had misreported whether they had taken any remedial mathematics in college. The greatest threat to the results would be low-propensity students claiming not to have taken remedial mathematics in college even though they actually did (Adelman, 1999; Kirst, 2007); however, even if 30% of the Prealgebra and Algebra I students did misreport and none of the high-propensity students misreported, the average effect size would still have only been tau = .10 among the 11 strata and tau = .04 among the lowest four strata.

Table 1

Estimated Effect of Highest Mathematics Class on Placement out of Postsecondary Remedial Mathematics, by Stratum

	Number of students							95%
						Covariate		error
Stratum	Prealgebra	Algebra I	Geometry	Algebra II	Precalculus	balance	Tau	margin
	54	50	48	45	14	98.1%	–0.05	0.18
2	19	45	66	61	20	98.8%	–0.12	0.17
3	18	39	65	69	21	98.8%	0.03	0.16
4	23	36	86	141	30	98.6%	0.02	0.14
5	18	27	83	137	53	98.9%	0.06	0.14
6	11	30	68	145	63	97.6%	0.06	0.13
7	4	20	63	151	80	90.4%	–0.04	0.13
8	5	13	56	171	72	91.8%	0.05	0.14
9	8	10	42	160	97	94.8%	0.13	0.12
10	5	3	40	156	113	79.1%	0.06	0.12
11	2	4	38	160	113	66.0%	0.12	0.11

Note. Prealgebra group includes students whose highest reported class was Prealgebra, General Math, or Business Math; Precalculus group includes students whose highest reported class was Precalculus, Trigonometry, or Calculus. Covariate balance is the percentage of pairwise comparisons on covariate means between the five treatment groups that are not significantly different. Effect size (tau) is weighted by marginal mean weights. Error margins account for the inflated variance due to the design effect. Effect sizes for top two strata may still include selection bias on the observed variables.

Thus, none of the averaged effects reach the minimum practical effect size of tau = .20 set forth by Ferguson (2009). In other words, these results suggest that mathematics coursetaking in the algebra-calculus pipeline does not have a practical effect on helping a student place out of remedial mathematics in college. Moreover, any effect that does exist is particularly weak for the students least likely to take an upper-level mathematics course.

Selection bias in a scatterplot. Because propensity scores and effect sizes can both be considered quantitative variables, it is possible to plot their relationship on a scatterplot. This allows us to visually check for an interaction between propensity score stratum and effect size. Simply put, do our results suggest that lower-propensity students receive a different degree of treatment effect than the higher-propensity students? This is a particularly relevant question for policymakers who, by mandating that all students take certain math classes in order to graduate, would mostly be affecting the students least likely to take the upper-level classes otherwise. The scatterplot is shown in Figure 7.

The leftmost 11 points in Figure 7 represent the effect sizes within the strata for the comparable students. These are the same effect sizes that were presented in Table 1, but the scatterplot allows us to see the interaction between effect size and propensity score stratum more clearly. Within this subsample of comparable students, we see that there was very little effect observed, particularly for the students least likely to enroll in upper-level mathematics courses. The points in the gray box at right are not based on actual data; the *ELS* lacks the counterfactual data needed to create reliable estimates for these effect sizes. Instead, these hypothetical effect sizes have been extrapolated based on the assumption that the trendline found for the first 11 strata continues throughout the other two thirds of the analytic sample.

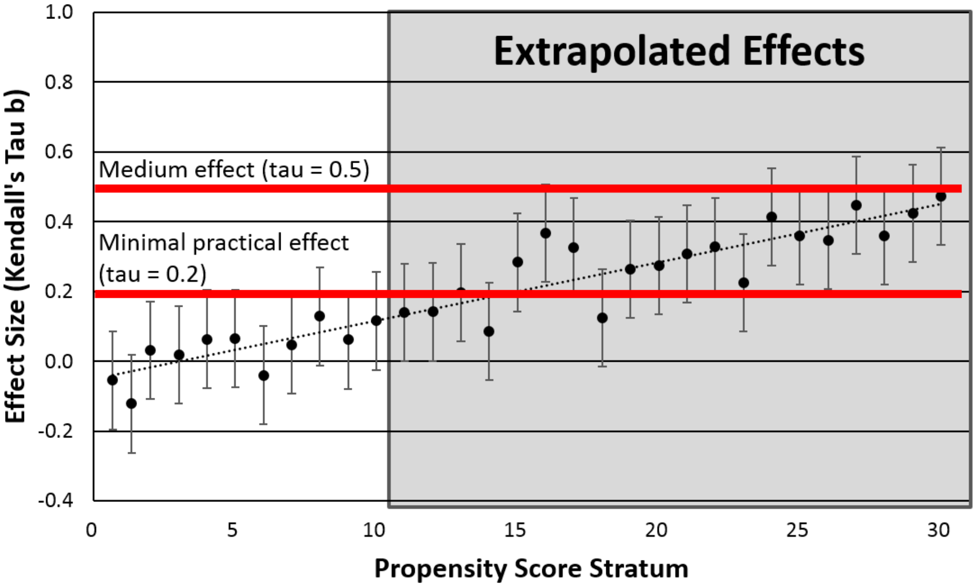

Figure 7. Estimated effect sizes for the entire analytic sample. Effect sizes for all strata in the gray box are extrapolated based on the linear trend in the strata from the sample of comparable students. In the strata based on available data, the effect size generally did not reach practical significance. In the strata based on extrapolation, most of the effect sizes did reach practical significance. Any actual results obtained in the extrapolated region would be at high risk for selection bias.

The word *extrapolation* should raise warning flags for those with even a modest background in statistics. There is no guarantee that the effect sizes for the students *without* common support were as shown in Figure 7; they could have been much smaller or much larger. The *ELS*, as well as any currently available nationally representative dataset, lacks the data needed to estimate effect sizes for these students. However, the absence of these important data is obscured when using methods such as regression that can produce a numerical result with or without common support (Stuart, 2010).

Moreover, even if the effect sizes did continue as extrapolated, a simple average of the estimated effect overlooks the interaction between effect size and propensity. Why is this interaction important? It tells us that students who were least likely to take advanced mathematics classes in high school benefitted the least from actually taking them (see the lower strata in Figure 6 and Figure 7). Similarly, students who were more likely to take advanced mathematics but did not, on average, missed out on an opportunity that might have been particularly beneficial to them (relative to their peers). Extrapolating effect sizes for students who lack counterfactual data leaves the door wide open to selection bias.

Generalizing the unbiased results back to the population. After calculating the estimated effect sizes within each propensity score stratum, we can generalize the results back to the population. Since we removed students who never attended college and those whose propensity scores were too high to be comparable to other students, these results apply to the third of college-bound sophomores from 2002 who were least likely to ever take Precalculus. It is important to note that these are precisely the students who are impacted by state mandates to take upper-level courses (the mandates won't have much impact on students who are likely to take the upper-level courses anyway). Do these results generate to the high school students of today, or of tomorrow? Not necessarily—system-wide changes have occurred in the past 15 years that may impact these results. But the results do update currently cited results on the topic which all (a) compared incomparable students, (b) were at high risk for heavy selection bias, or (c) were only conducted on the students within a particular state.

In this final step of generalizing results to the population, observational datasets such as *ELS* enjoy a major advantage over most true experiments. Because the sample was chosen randomly from the population rather than at convenience, the findings can be applied confidently to the population. In other words, a true experiment produces unbiased results for the sample, but invokes a margin of error when extended to the population; causal inferences made from randomly sampled observational data yield results with a margin of error but can be extended back to the population in an unbiased manner.

Concluding Remarks

Why does it matter that selection bias may be present in many results in the broader educational context? The disheartening racial and socioeconomic inequities published in the Coleman Report (Coleman et al., 1966) half a century ago are just as present in the education system of today (ACT, 2007; Gutstein, 2006; Schoenfeld, 2002). These inequities, combined with the prohibitive costs and ethics of conducting randomized control trials, place educational studies at particularly high risk for contamination from selection bias. So what? This means that we often assume that what works for the majority of students works for all students—a shaky claim when we are designing programs for those who are struggling most, either financially or academically. Even worse, we deem treatments as effective that really did not have a practically significant effect for anyone. This can result in pouring large amounts of time and energy into ineffective solutions; time and energy that could instead be invested more meaningfully.

Though several options exist to help filter out selection bias from observational data, most of them can be quite intimidating for the casual researcher. Yet, there is hope! A researcher could partner with someone who is knowledgeable about causal inference methods. Or, if such a partnership was not practical, the researcher could at least identify important variables, use software to impute the missing values, and then run a logistic regression with the treatment as the dependent variable. This would produce good enough propensity score estimates to allow the researcher to get a feel for how much selection bias could impact the results—and then report this risk accurately, even if he or she did not plan to remove it. But most importantly, educators should simply be aware of the constant risk of selection bias to be able to use a sniff test when they read studies where (a) causal effects are claimed on observational data without mentioning how selection bias was addressed or (b) too much weight is given to correlational results even though it is fairly clear that the students in one treatment group are likely quite different from those in another treatment group (or the control).

The mere fact that treatment groups are incomparable does not necessarily mean that selection bias will change the significance or the effect size of a treatment. Negligible selection bias is reasonable to assume when correlational studies and causal studies on a topic yield similar results. Differing results, on the other hand, are suggestive of the influence of selection bias on the correlational results, although competing explanations include issues of data quality or methodological accuracy.

Educational decisions are often dependent on an assumption that Treatment A (e.g., a law, an intervention, funding) will increase the likelihood of Outcome B. And, if the most relevant studies on a particular topic only show an association, the decision maker either embraces the association as causation or is left uncertain about the validity of the results as they apply to a causal inference. This is not to imply that correlational research is useless or that researchers working with observational data should always attempt to make a causal claim—this wouldn’t be practical for small scale studies like action research, nor is it necessary for making low stakes decisions that do not substantially impact students’ lives or require major resource allocation. However, researchers working with observational data should be expected to account for or at least identify the presence of selection bias to the degree that the data permit it. At a minimum, every educational researcher should become informed about the concerns surrounding selection bias and explicitly warn the reader of selection bias when it hasn’t been sufficiently accounted for. And every reader of educational research should learn to have a healthy skepticism when reading observational studies that made no attempt to account for selection bias.

Author Notes

Daniel A. Showalter is an Assistant Professor of Mathematics at Eastern Mennonite University in Harrisonburg, VA.

Luke B. Mullet is a mathematics major at Eastern Mennonite University in Harrisonburg, VA.

Correspondence concerning this article should be addressed to Daniel Showalter at Daniel.showalter@emu.edu.

References

Achieve. (2008). *The building blocks of success: Higher-level math for all students*. Retrieved from Achieve website: http://www.achieve.org

ACT. (2007). *Rigor at risk: Reaffirming quality in the high school core curriculum*. Retrieved from ACT website: http://www.act.org

Adelman, C. (1999). *Answers in the toolbox: Academic intensity, attendance patterns, and bachelor's degree attainment*. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education. Retrieved from http://www.ed.gov/pubs/Toolbox/toolbox.html

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: U.S. Department of Education. Retrieved from http://www.eric.ed.gov/PDFS/ED490195.pdf

Allison, P. D. (2002). *Missing data [Series: Quantitative applications in the social sciences 136]*.Thousand Oaks, CA: Sage Publications.

Altonji, J. G. (1995). The effects of high school curriculum on education and labor market outcomes. *The Journal of Human Resources, 30*(3), 409–438.

Aughinbaugh, A. (2012). The effects of high school math curriculum on college attendance: Evidence from the NLSY97. *Economics of Education Review, 31*, 861–870.

Bailey, T. (2009). *Rethinking developmental education in community college (Brief No. 40)*. New York, NY: Columbia University, Teachers College, Community College Research Center.

Barber, R. (2011). *Characteristics of students placed in college remedial mathematics: Using the ELS 2002/2006 data to understand remedial mathematics placement (Unpublished doctoral dissertation)*. Arizona State University, Tempe, AZ.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Currie, J. (2003, June). *When do we really know what we think we know? Determining causality*. Invited paper presented at Work, Family, Health, and Well-Being conference, NICHD Administration for Children and Families, Washington D.C.

Deil-Amen, R., & Rosenbaum, J. E. (2002). The unintended consequences of stigma-free remediation. *Sociology of Education*, 249–268.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538.

Goodman, J. (2012). *The labor of division: Returns to compulsory math coursework* (HKS Faculty Research Working Paper Series RWP12-032). Retrieved from John F. Kennedy School of Government, Harvard University website: http://nrs.harvard.edu/urn-3:HUL.InstRepos:9403178

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Gutstein, E. (2006). *Reading and writing the world with mathematics: Toward a pedagogy for social justice*. New York, NY: Routledge.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Ho, D., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

Hong, G. (2004). *Causal inference for multi-level observational data with application to kindergarten retention* (Doctoral dissertation). The University of Michigan, Ann Arbor, MI.

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multi-level data. *Journal of Educational and Behavioral Statistics*, *35*, 499 –531. doi:10.3102/1076998609359785

Hong, G. (2012). Marginal mean weighting through stratification: A generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychological Methods*, *17*(1), 44–60.

Hoyt, J. E. (1999). Level of math preparation in high school and its impact on remedial placement at an urban state college. *College and University*, *74*(3), 37–43.

Hoyt, J. E., & Sorenson, C. T. (2001). High school preparation, placement testing, and college remediation. *Journal of Developmental Education*, *25*(2), 26–34.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 171, 481–502.

Kim, J., Kim, J., DesJardins, S. L., & McCall, B. P. (2015). Completing Algebra II in high school: Does it increase college access and success? T*he Journal of Higher Education*, *86*(4), 628–662.

Kirst, M. (2007). Who needs it? Identifying the proportion of students who require postsecondary remedial education is virtually impossible. *National CrossTalk*, *15*, 11–12.

Lechner, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society*. *Series A (Statistics in Society)*, *165*(1), 59–82.

Leigh, J. P., & Schembri, M. (2004). Instrumental variables technique: Cigarette price provided better estimate of effects of smoking on SF-12. *Journal of Clinical Epidemiology, 57*(3), 284–293.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University Press.

Long, M. C., Iatarola, P., & Conger, D. (2009). Explaining gaps in readiness for college-level math: The role of high school courses. *Education Finance and Policy*, *4*(1), 1–33.

Loveless, T. (2013). *How well are American students learning? The 2013 Brown Center Report on American Education*. Washington, DC: Brookings Institution Press.

National Education Association. (1894). *Report of the Committee of Ten on secondary studies with the report of conferences arranged by the committee*. New York: American Book Company.

Reys, R., & Reys, R. (2009). Two high school-mathematics curricular paths: Which one to take? *Mathematics Teacher*, *102*, 568–570.

Rose, H., & Betts, J. R. (2004). The effect of high school courses on earnings. *The Review of Economics and Statistics*, *86*(2), 497–513.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Roth, J., Crans, G. G., Carter, R. L., Ariet, M., & Resnick, M. B. (2001). Effect of high school course-taking and grades on passing a college placement test. *High School Journal*, *84*(2), 72–84.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects: Using experimental and observational designs*. Washington, DC: American Educational Research Association.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing, and equity. *Educational Researcher*, *31*(1), 13–25.

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2012). *Improving the targeting of treatment: Evidence from college remediation* (No. w 18457). National Bureau of Economic Research.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334–1356.

St. John, E. P., & Chung, A. S. (2006). Access to Pure Math. In E. P. St. John (Ed.) *Education and the public interest: School reform, public finance, and access to higher education* (pp. 135–162). New York: Springer.

Strong American Schools. (2008, September). *Diploma to nowhere*. Retrieved from http://www.voorheesgroup.org

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of "A critical appraisal of propensity score matching in the medical literature between 1996 and 2003," *Statistics in Medicine*, *27*(12), 2062–2065.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci.*, *25*(1), 1–21. doi: 10.1214/09-STS313

U.S. Department of Education (U.S. DOE), National Center for Education Statistics. (2004). *Education Longitudinal Study of 2002: Base year data file user's manual*, NCES 2004-405, by Steven J. Ingels, Daniel J. Pratt, James E. Rogers, Peter H. Siegel, and Ellen S. Stutts. Project Officer: Jeffrey A. Owings. Washington, DC: Author.

What Works Clearinghouse. (2013). *Procedures and standards handbook: Version 3.0*. Washington, DC: Author.

Appendix

Table A1
Weighted Covariate Balance by Propensity Score Stratum

Stratum	Nonsignificant pairwise differences (%)	Students	Minimum propensity score	Maximum propensity score
1	98.1	211	.01	.03
2	98.8	211	.03	.05
3	98.8	212	.05	.08
4	98.6	316	.08	.11
5	98.9	318	.11	.15
6	97.6	317	.15	.19
7	90.4	318	.19	.24
8	91.8	317	.24	.28
9	94.8	317	.28	.32
10	79.1	317	.32	.37
11	66.0	317	.37	.42

Note. Within each stratum, Games-Howell pairwise difference tests were conducted between every two groups on every covariate. All data were weighted with marginal mean weights before conducting significance tests. Random assignment would, on average, result in 95% of the pairwise differences being nonsignificant; achieving roughly this same number with our propensity scores would reassure us that we have successfully removed the selection bias, at least in terms of the observed covariates. As seen in the second column, we achieved this goal in all but the final two strata; there is still risk for selection bias in these two strata.

Table A2

Standardized Differences in Means and Variances of Influential Predictors Before and After Applying Marginal Mean Weights

		Cohen's *d* (*M*)		*SD*	Cohen's *d* (*Var*)	
Predictor	*SD*	Before	After	(quadratic)	Before	After
Math IRT score	11.15	0.79	0.17	907.62	0.72	0.14
Math score, standardized test	9.22	0.79	0.17	981.58	0.73	0.15
Math teacher's expectations	1.22	0.88	0.21	10.50	0.80	0.18
Math proficiency, level 4	0.37	0.49	0.09	0.35	0.41	0.07
Western region	0.39	0.17	0.26	0.39	0.17	0.26
Recommended for AP/honors class	0.39	0.39	0.07	0.38	0.35	0.03
Reading IRT score	9.10	0.69	0.25	568.59	0.63	0.21
Math self-efficacy	0.90	0.32	0.25	1.06	0.13	0.20
English teacher's expectations	1.23	0.83	0.20	10.84	0.76	0.17
Reading proficiency, level 2	0.38	0.64	0.22	0.40	0.57	0.19
Math proficiency, level 5	0.09	0.13	0.00	0.07	0.09	0.00
Uses graphing calculators in math class	1.65	0.26	0.20	10.11	0.26	0.18
Writing ability	0.90	0.65	0.19	1.09	0.18	0.18
Midwest region	0.44	0.09	0.27	0.44	0.09	0.27
Southern region	0.48	0.25	0.36	0.48	0.25	0.36
% 10th graders on college prep path	30.97	0.32	0.31	3636.26	0.32	0.30
Asian/Pacific Islander	0.31	0.15	0.23	0.31	0.15	0.23
% 10th graders receive remedial math	7.89	0.26	0.29	294.49	0.20	0.29
Type of school student plans to attend	0.43	0.67	0.30	1.58	0.61	0.29
Thinks math is fun	0.77	0.23	0.22	4.11	0.22	0.24

Note. Effect sizes shown are the standardized averages of all pairwise differences between treatment groups. Standardized averages after the application of marginal mean weights are calculated based on a weighted mean of average differences within each stratum. The SD used to standardize the differences in means and the SD used to standardize the differences in variances are both taken from the analytic sample and used for all comparisons as suggested by Stuart (2008). Quadratic terms were formed by squaring each data point for each variable. The small values of Cohen's *d* overall suggest that selection bias has been reduced to an acceptable level on these influential variables.